

O'REILLY®

# Fundamentals of Large Language Models





# Week 1 (today)

## GPT-3 and why we needed GPT-4

- Where GenAI stands now
- GPT-3+ and GPT-4
- Limitations of GPT-4
- RLHF
- Q&A
- Break

## Transfer learning and fine-tuning

- BERT
- Transfer learning and fine-tuning
- Q&A
- Break

## Transformer architecture overview

- Encoders and decoders
- Q&A
- Break



# Week 2 (next week)

## LLM embeddings lab

- What are they?
- Exercises and demos
- Q&A
- Break

## Benchmarks

- LLMs - HELM
- Q&A
- Break

## LLMs 1-2 years after GPT-3

- Scaling laws Chinchilla
- BIG-Bench
- PaLM
- OPT and BLOOM and Llama2
- Mistral
- Q&A



**This online training is always being updated.**

**Will look at Llama-3.1**

**Previous versions:**

- Falcon -> Llama 2 -> Mistral 7B -> Mixtral-8x22b**
- GPT-4 -> GPT-4-turbo -> GPT-4o**



Live Course



## Image Generation using Stable Diffusion and Midjourney

With [Jonathan Fernandes](#)

⌚ 4h 0m   🗓 Aug 15 • 5pm-9pm

Live Course

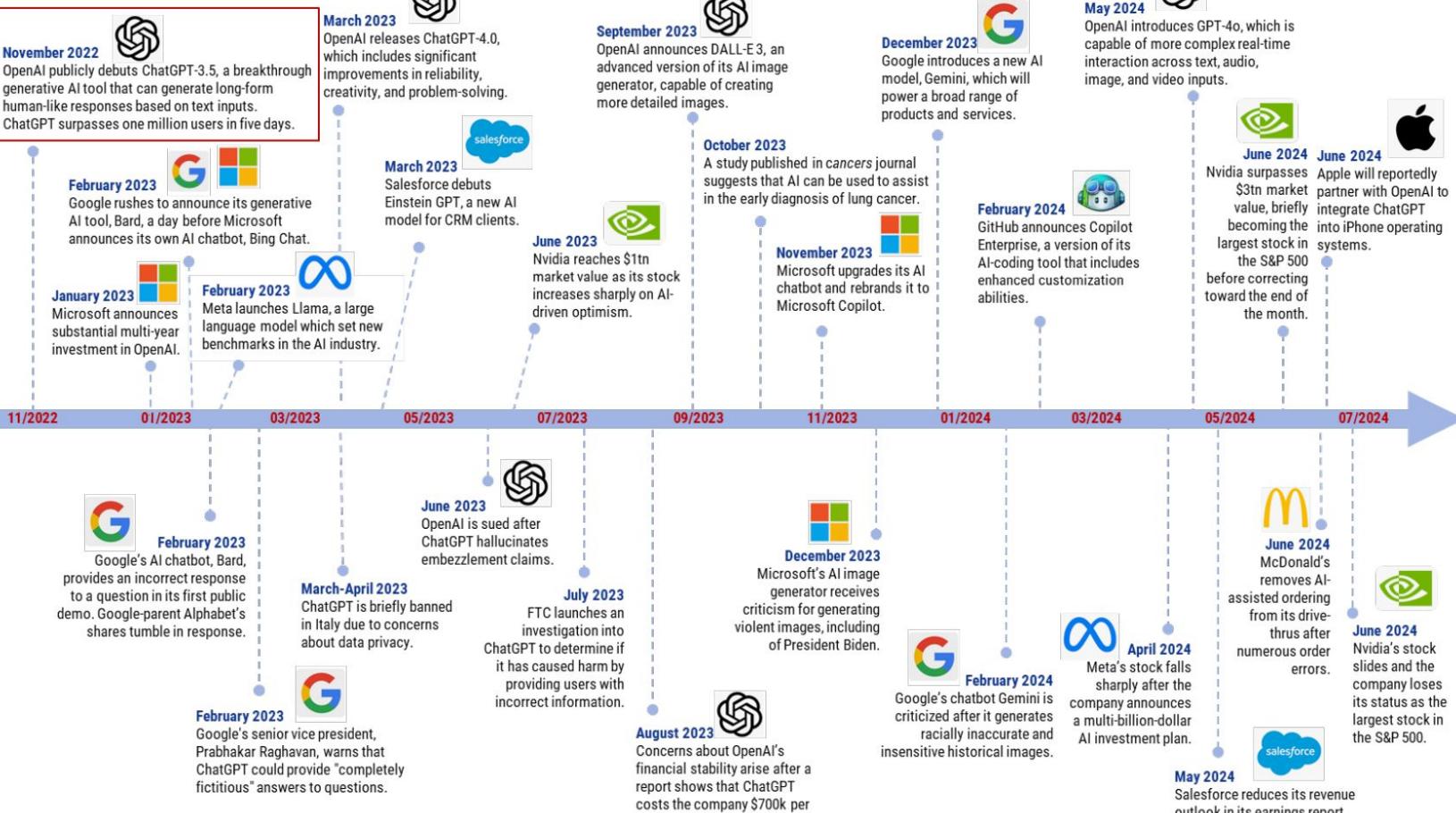


## Hands-on Retrieval Augmented Generation (RAG)

With [Jonathan Fernandes](#)

⌚ 3h 0m   🗓 Aug 29 • 5pm-8pm

# AI triumphs...



[https://www.goldmansachs.com/images/migrated/insights/pages/gs-research/gs-research.pdf-too-much-spend-too-little-benefit-TOM\\_AI%202.0\\_ForRedaction.pdf](https://www.goldmansachs.com/images/migrated/insights/pages/gs-research/gs-research.pdf-too-much-spend-too-little-benefit-TOM_AI%202.0_ForRedaction.pdf)

...and disappointments



# Why Zuckerberg's multibillion-dollar gamble doesn't just matter to Meta

*Dan Milmo and Alex Hern*

As Llama 3.1 405B is made freely available, investors are asking when the huge industry spend will pay off





# Why Zuckerberg's multibillion-dollar gamble doesn't just matter to Meta

*Dan Milmo and Alex Hern*

As Llama 3.1 405B is made freely available, investors are asking when the huge industry spend will pay off



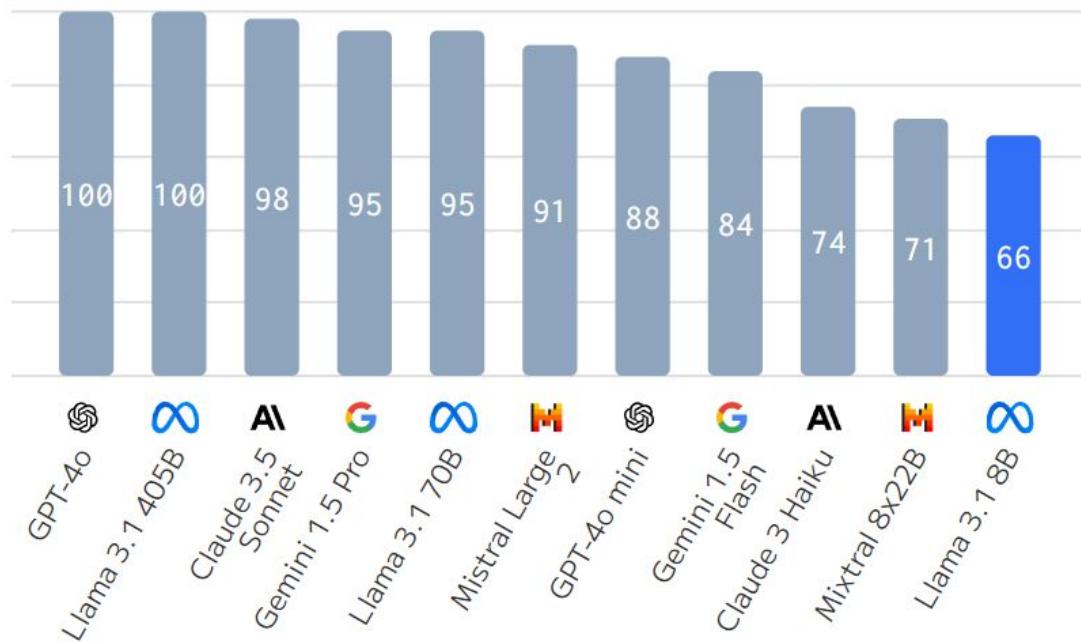
**Why would he make it available for free?**



# Quality

## QUALITY

*Quality Index; Higher is better*

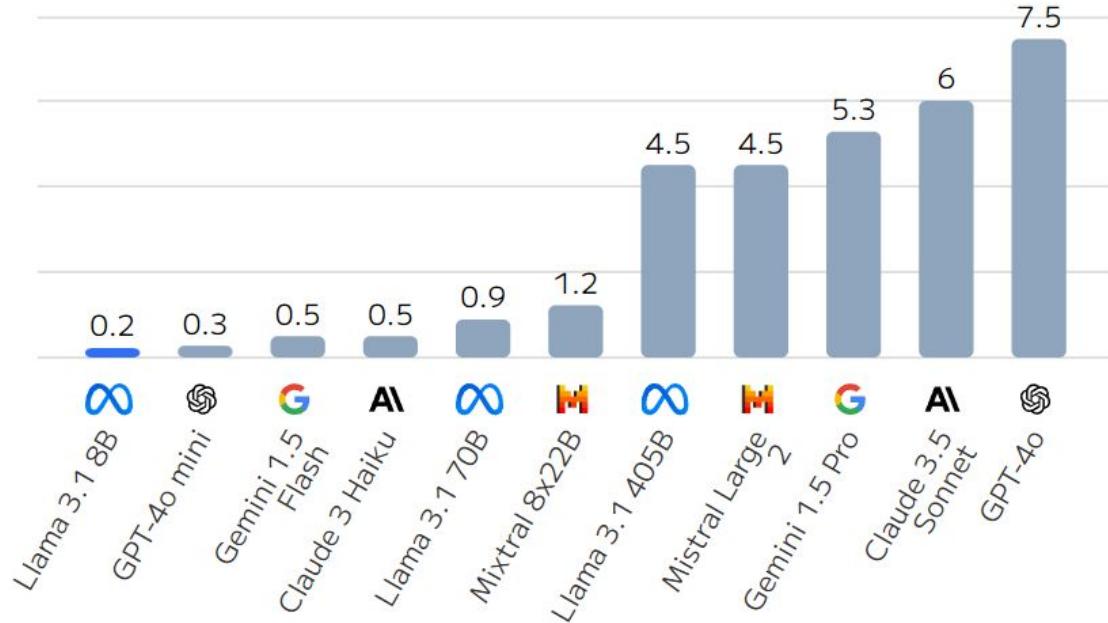




# Pricing

## PRICE

USD per 1M Tokens; Lower is better





TECH · A.I.

# Generative AI is getting kicked off its pedestal — it will be painful but it's not a bad thing

BY SHARON GOLDMAN

August 6, 2024 at 8:16 PM GMT+1



OpenAI CEO Sam Altman

<https://fortune.com/2024/08/06/generative-ai-reality-check-tech-selloff-new-phase-less-hype-more-roi/>



ARTIFICIAL INTELLIGENCE

# Will the \$1 trillion of generative AI investment pay off?

August 5, 2024   Share ↗



<https://www.goldmansachs.com/insights/articles/will-the-1-trillion-of-generative-ai-investment-pay-off>



Our perspective right now is that we're putting all this infrastructure in place to run these things. And we're seeing incredible improvements in how these models perform and what they can do. But, as Sung mentioned, we really need to see, at some point over the next year to year-and-a-half, applications that use this technology in a way that's more profound than coding and customer service chatbots.

If this ends up just doing coding and customer service, we're massively overspending on this. Again, I think both of us are very convinced that, over the intermediate term, we're going to see those applications and those use cases. And it's going to profoundly change how all of us do our work.

But I think the whole market is trying to figure out what else needs to take place for new applications and use cases to develop, and what you will see coming out the other side. So we're in that period right now where we need to see progress on that.



**Sung Cho:** We know there's not going to be more than four. There's nobody else that can compete. The only companies that can make this level of investment are Meta, Google, OpenAI, and Anthropic.

**Brook Dane:** But what we don't know yet is: As these models mature, and as you stop seeing these step-function increases, which will happen at some point, will the best model three years from now be so much better than everybody else's model that it takes 80% of the market share? Or do we have four really good models, and people will use them for different use cases, in different areas, and in different ways? Will we have four models at scale?

If there are four equally robust models, you would think that gets commoditized pretty quickly. Whereas if one of them becomes the clear, dominant leader, then it's going to have incredible economics. We don't know that answer yet.

But what we do know is that none of the four can afford to fall off the pace of innovation. Because if you do, if you stop at this first-year-college-student level of intelligence, and the other guys become Phd-level intelligence, it may be hard for you to have a market if they're that much better, and they're riding the efficiency and the cost curves down faster than you are.



# About me



**Jonathan Fernandes**  Get verified

Generative AI | Large Language Models | NLP

United Kingdom · [Contact info](#)



University of Warwick -  
Warwick Business School



## What does GPT stand for?

-  **Generative Pre-trained Transformer**
-  **General Pre-trained Transformer**



## What are the parameters for a Large Language Model?

-  **The size of the model**
-  **The variables that get adjusted during the training**



## What are tokens for a Large Language Model?

👍 These are subwords

👎 This is another word for parameter



## What is the size of the GPT-4 model?

175B

This information wasn't released



## How is GPT-4 different from GPT-3 and GPT-3.5?

-  **GPT-4 is multi-modal**
-  **GPT-4 is a decoder-based model**

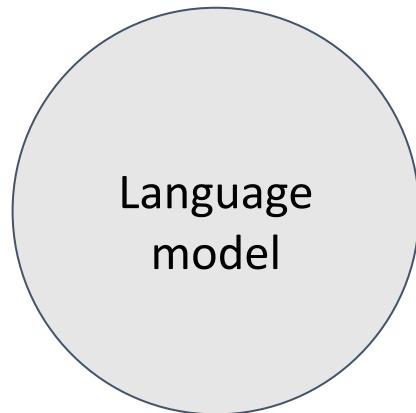


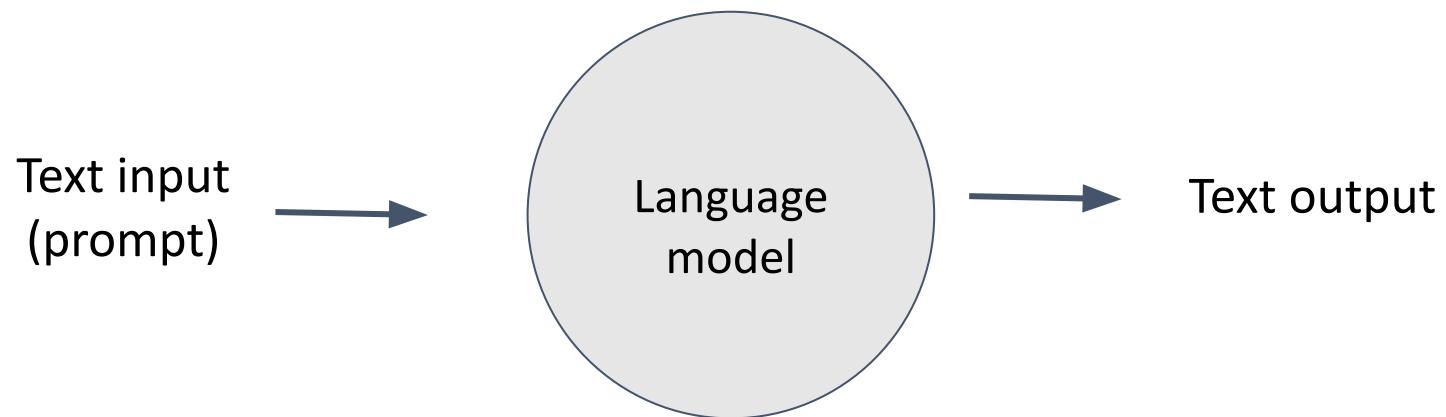
## How can you ensure that GPT-4 won't hallucinate?

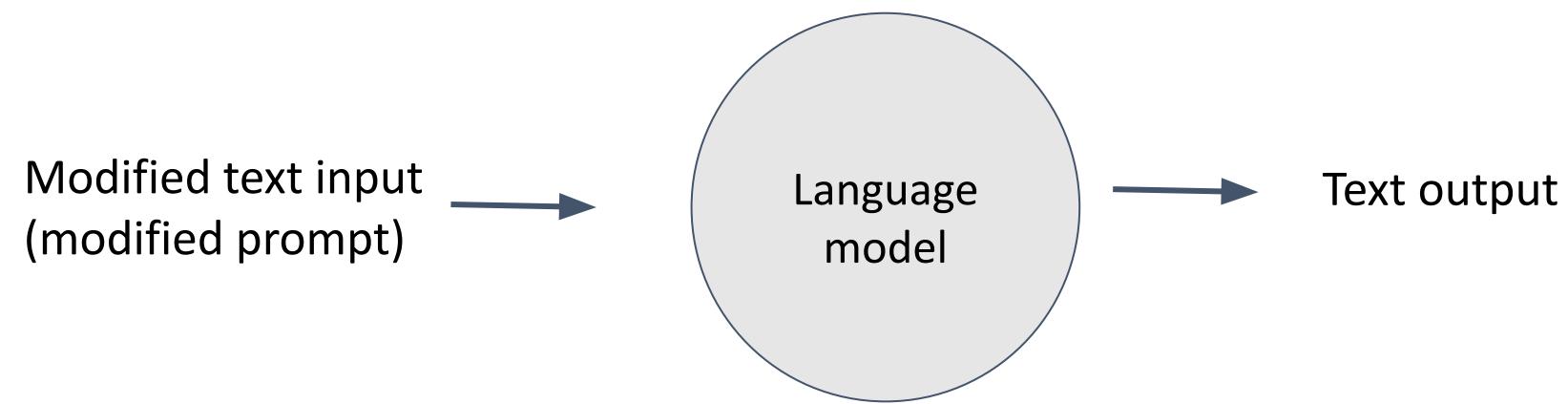
-  **Give it more training data**
-  **No way to do this at the moment.**

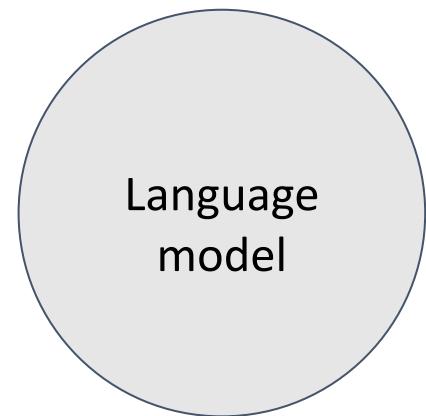
# What Are Large Language Models and GPT?

Text input  
(prompt)



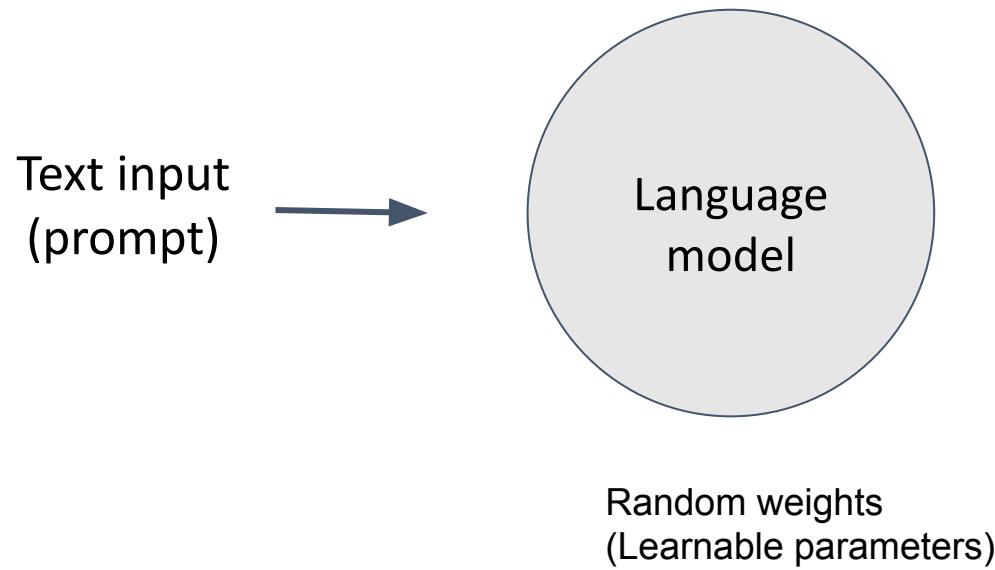




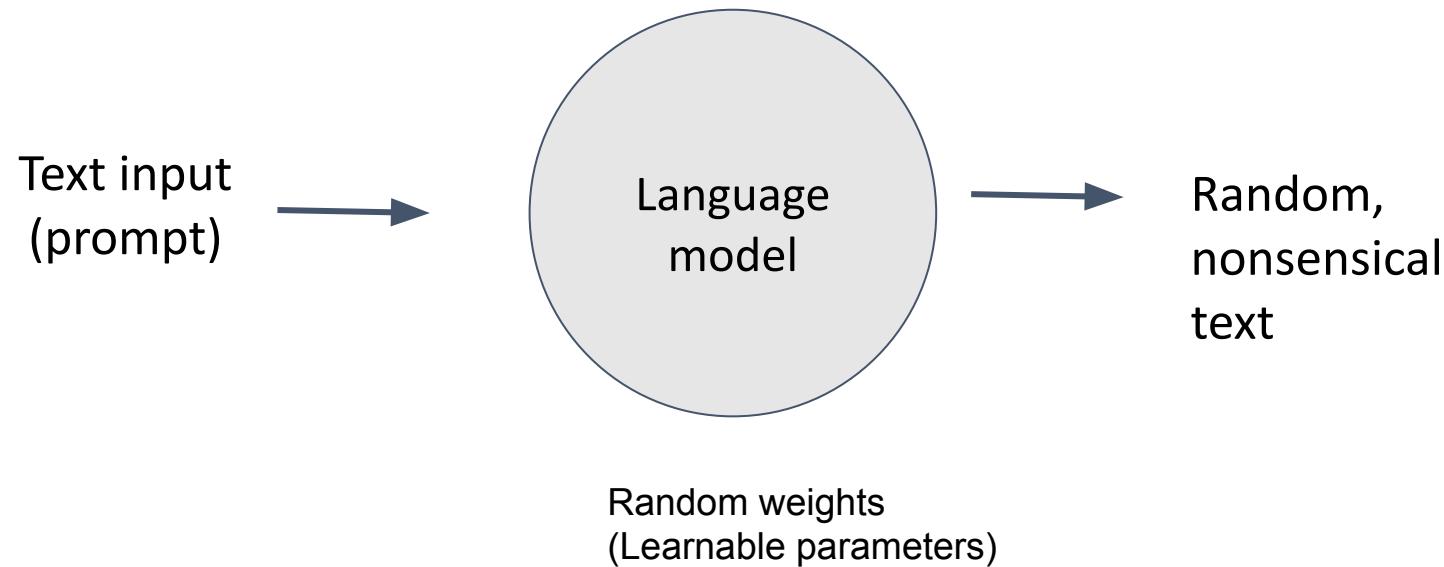


Parameters or  
Learnable parameters

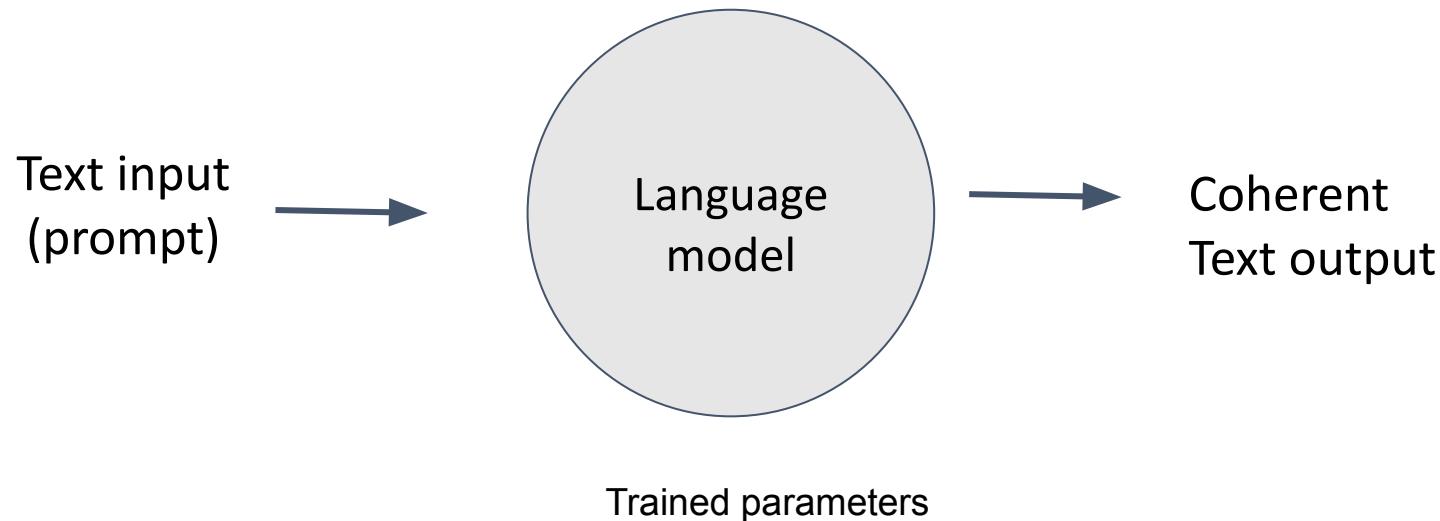
## Initial model (Before training Language model)



## Initial model (Before training Language model)

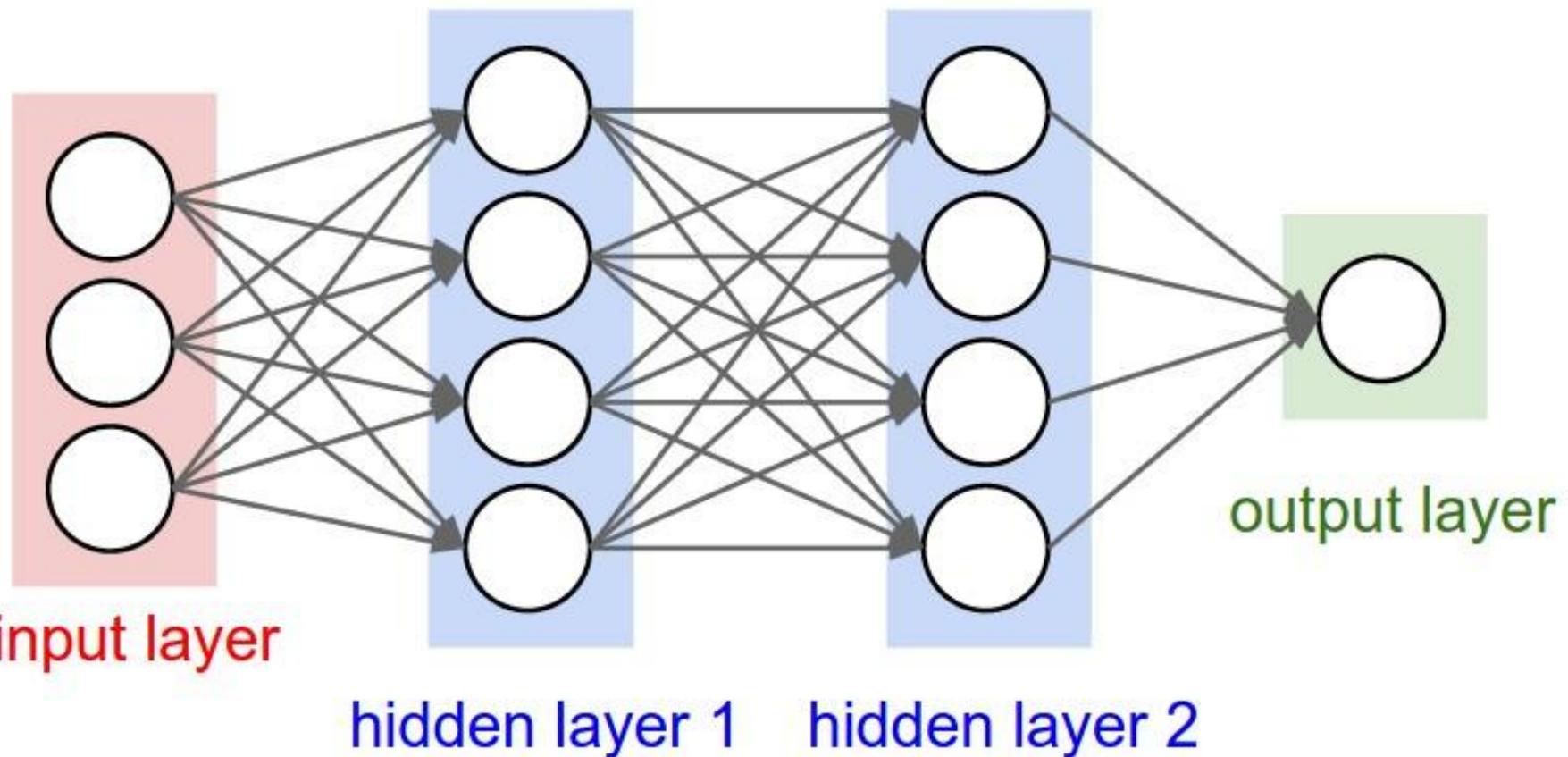


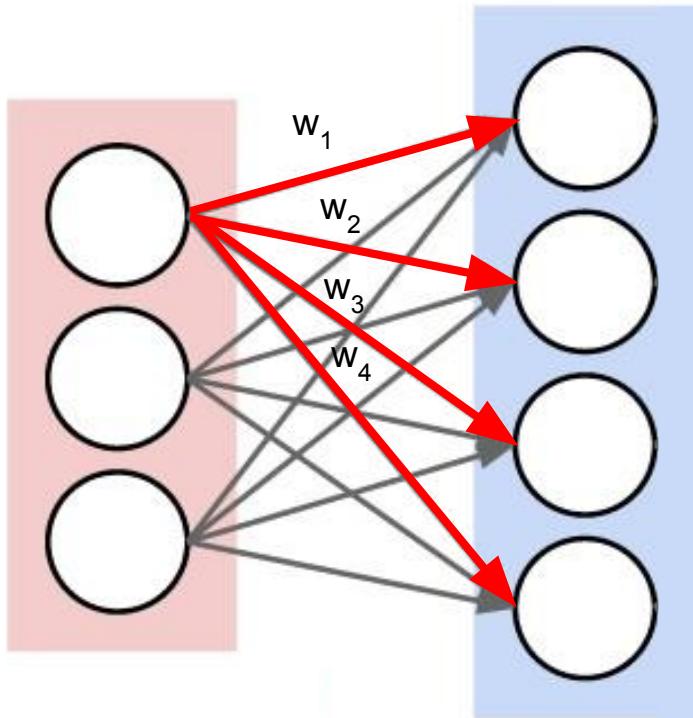
After training



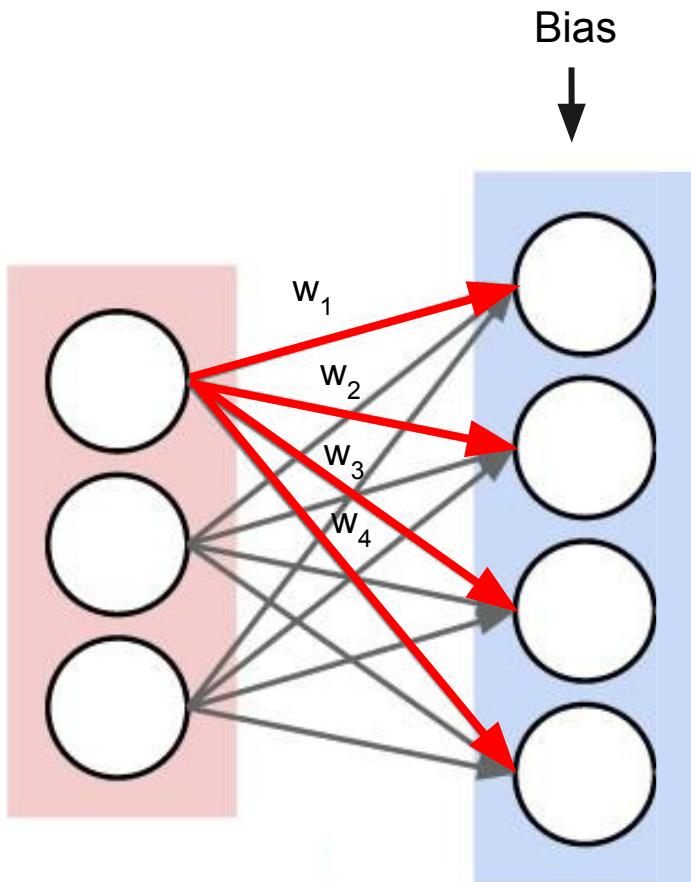


What are these (learnable) parameters?

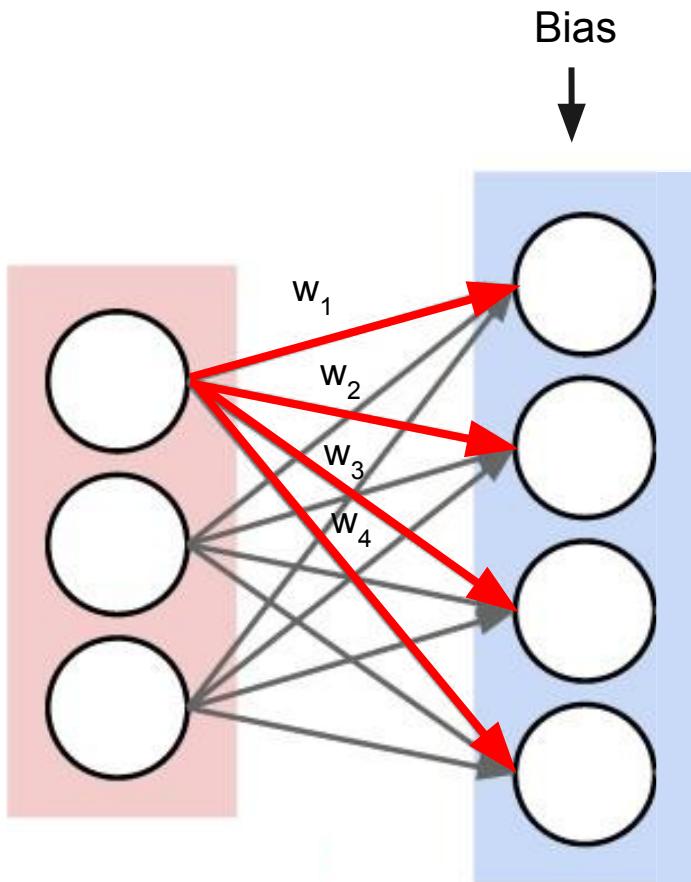




Weights:  
Inputs x Outputs

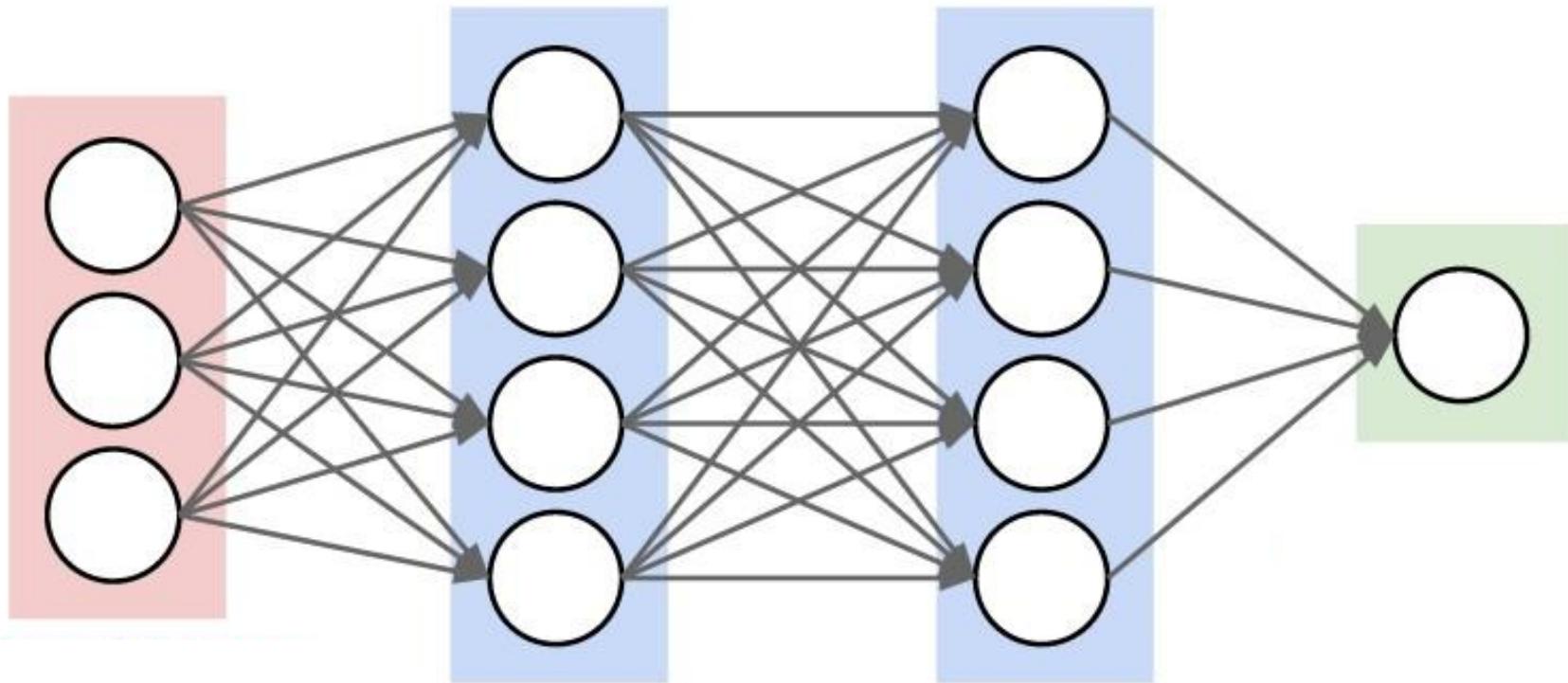


Bias



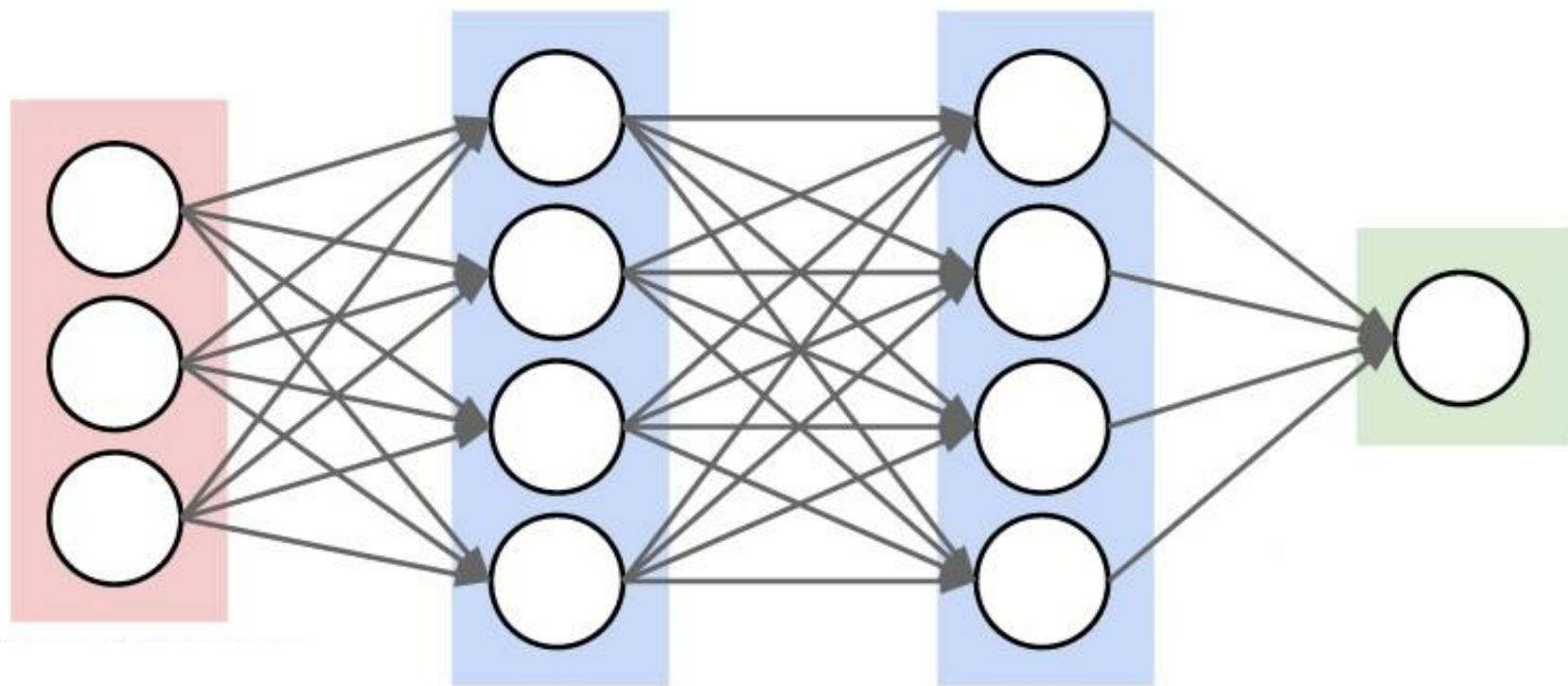
Number of parameters:  
Number of weights + Biases  
 $(3 \times 4) + 4$

$3 \times 4 + 4$   
**16**



$3 \times 4 + 4$   
**16**

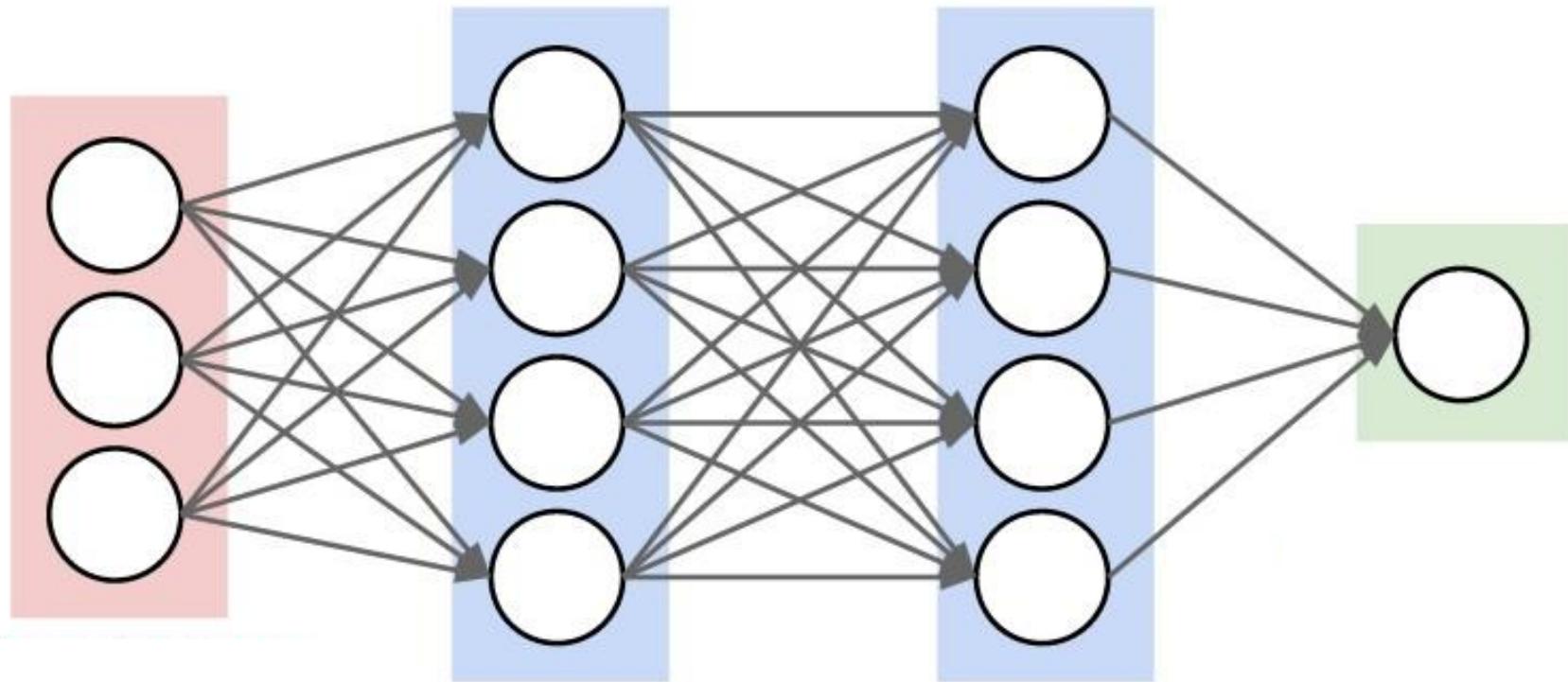
$4 \times 4 + 4$   
**20**



$3 \times 4 + 4$   
**16**

$4 \times 4 + 4$   
**20**

$4 \times 1 + 1$   
**5**

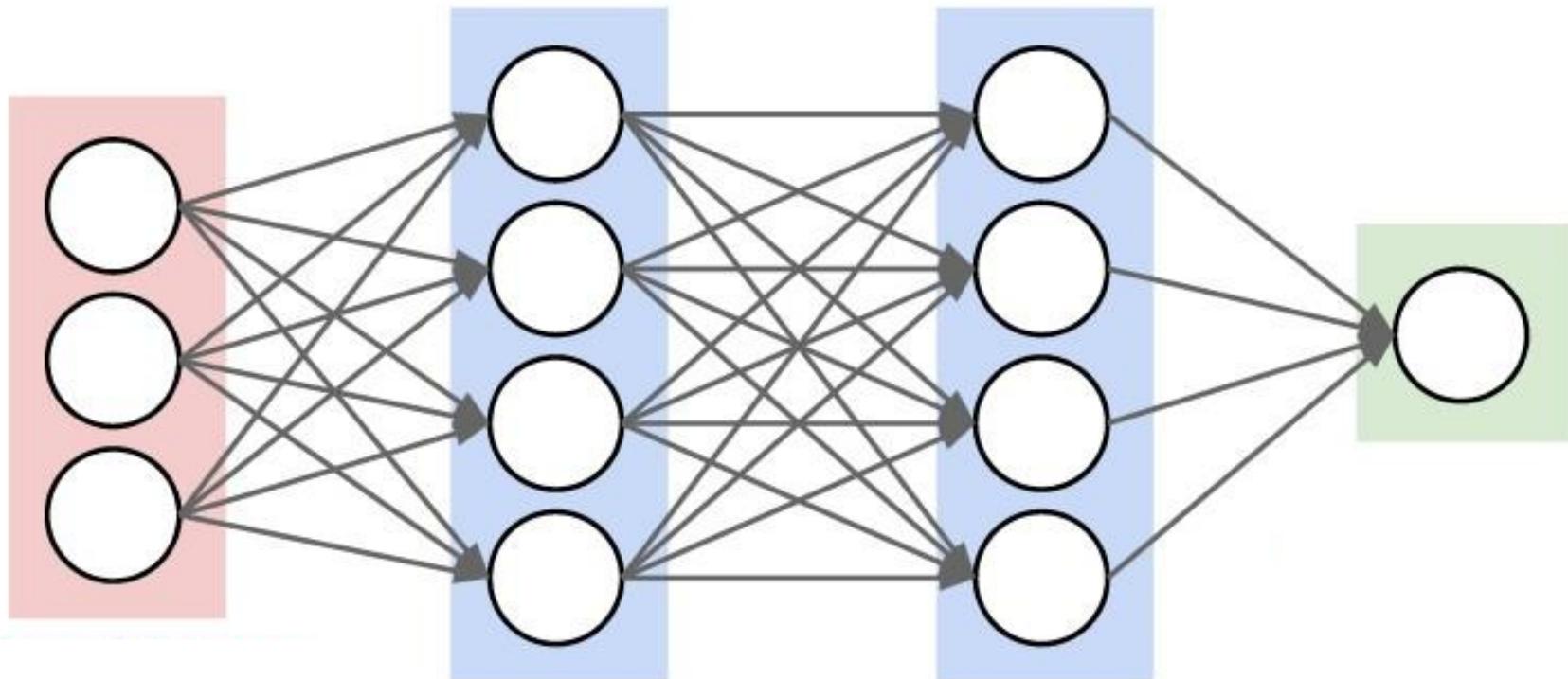


$3 \times 4 + 4$   
**16**

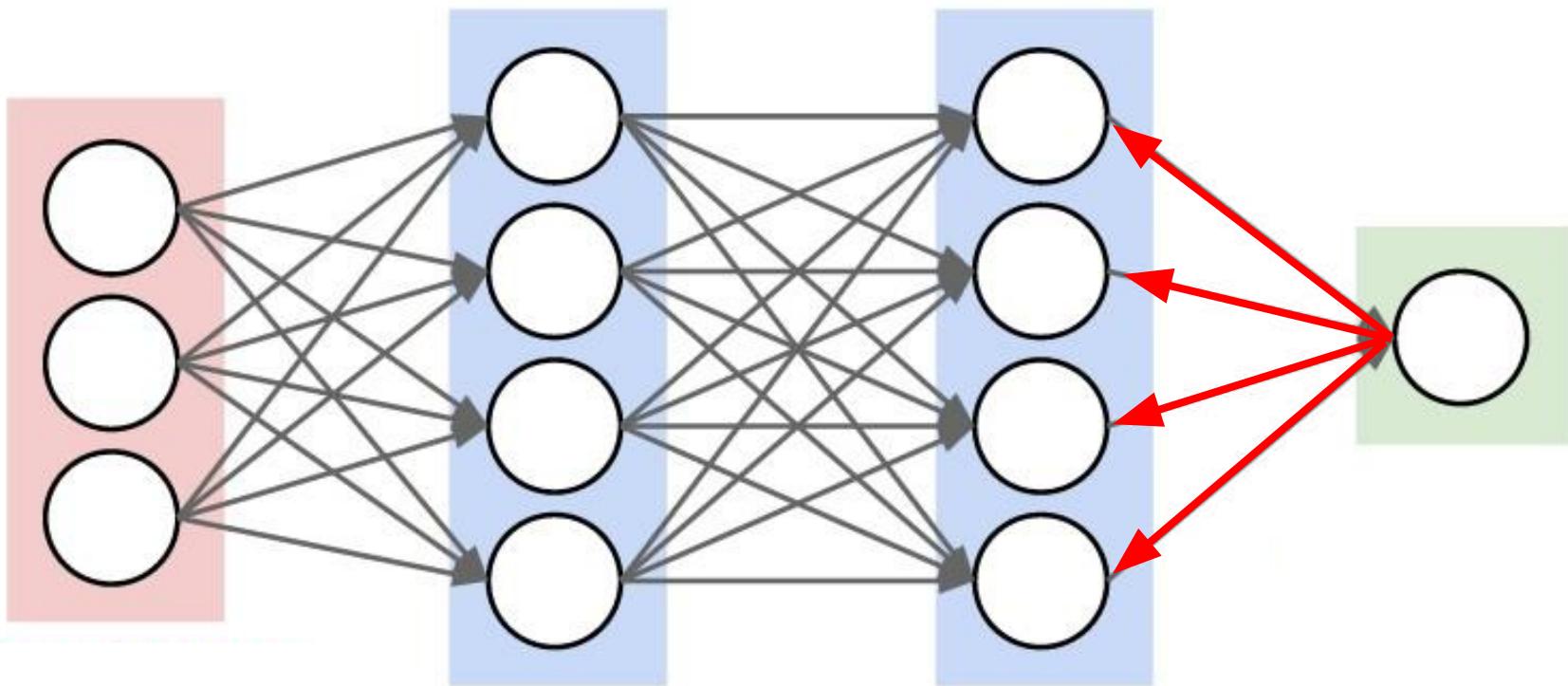
$4 \times 4 + 4$   
**20**

$4 \times 1 + 1$   
**5**

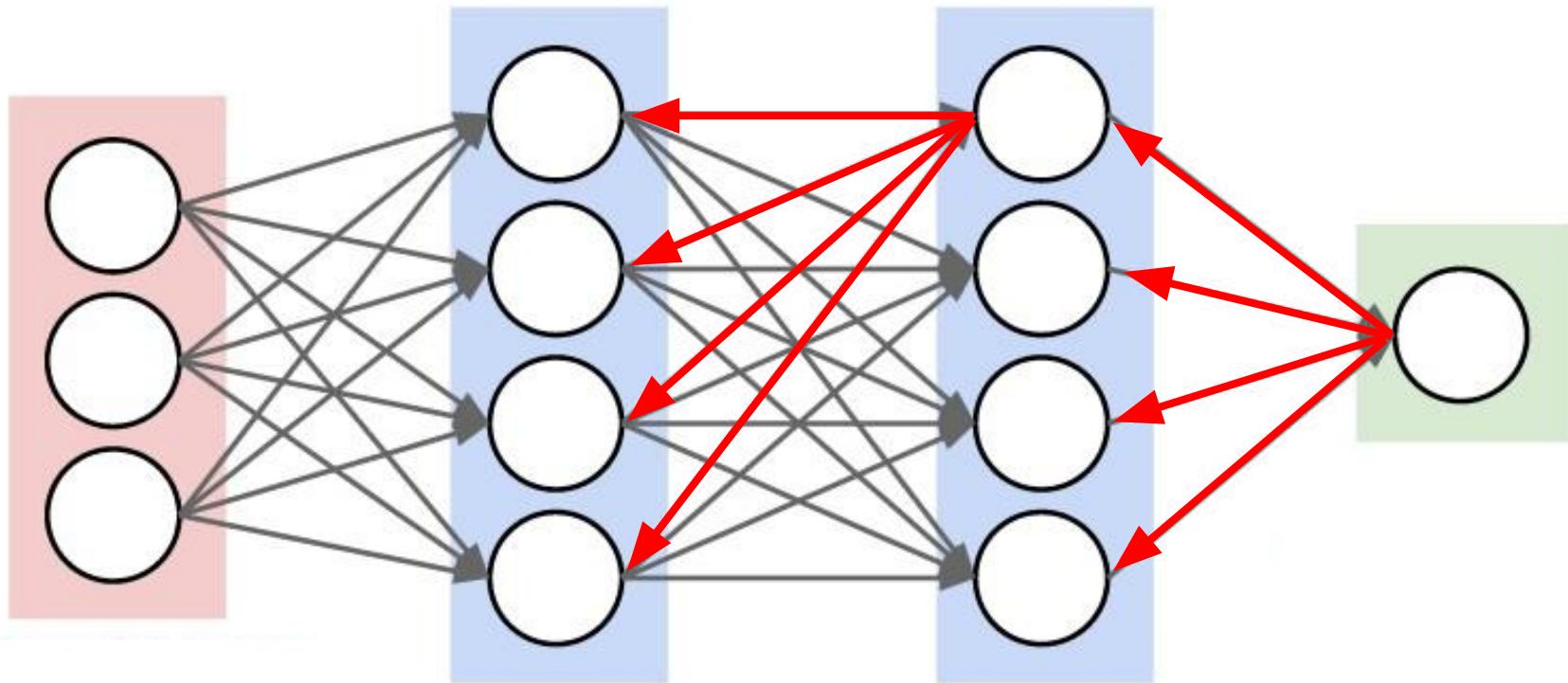
Total parameters:  
**41**



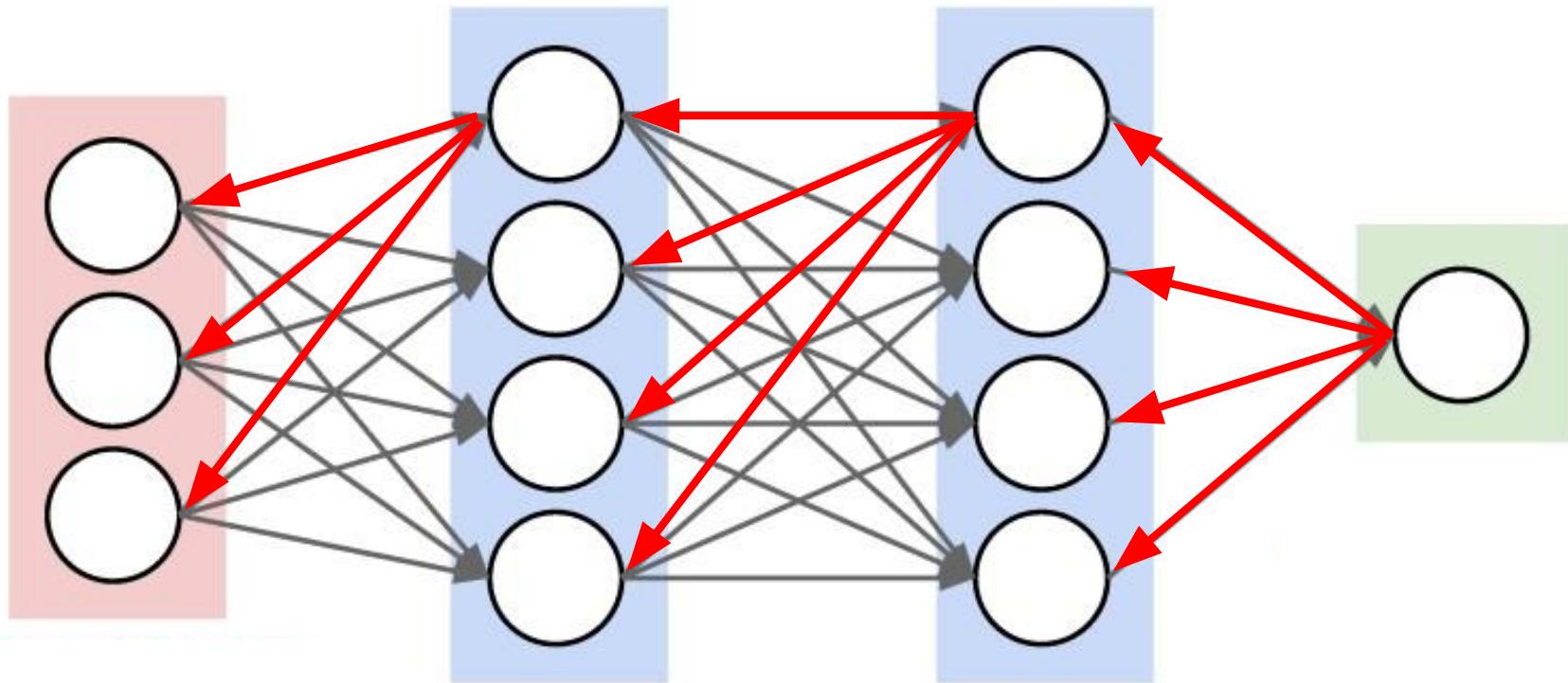
Backpropagation



Backpropagation



Backpropagation





# What are tokens?



# What are tokens?

GPT-3 Codex

Tokenization is the process of splitting words into smaller chunks or tokens. These tokens are then converted into token ids. These are numbers that are inputted into a language model.



[Clear](#) [Show example](#)

Tokens	Characters
37	185

Tokenization is the process of splitting words into smaller chunks or tokens. These tokens are then converted into token ids. These are numbers that are inputted into a language model.

[TEXT](#) [TOKEN IDS](#)

**G**enerative  
**P**re-trained  
**T**ransformer

**G**enerative – predicting a future token,  
given past tokens

**P**re-trained  
**T**ransformer

**G**enerative – predicting a future token,  
given past tokens

**P**re-trained – trained on a large corpus of  
data

**T**ransformer

**G**enerative – predicting a future token,  
given past tokens

**P**re-trained – trained on a large corpus of  
data

**T**ransformer – portion of transformer  
architecture

# Objectives of GPT-3

- Predict the next token, given preceding tokens
- Causal language models
- Autoregressive language models

Roses \_

Roses are \_

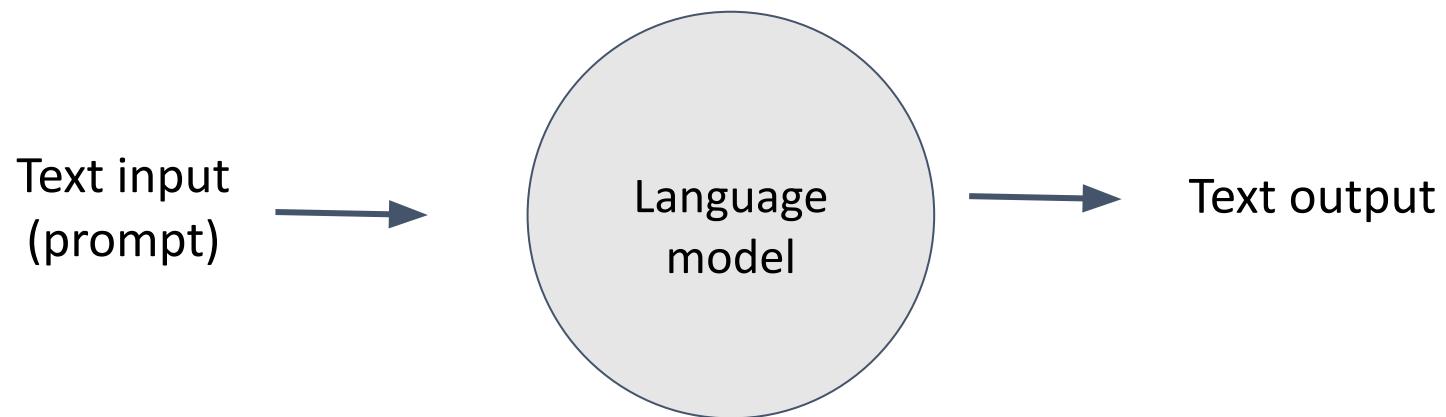
Roses are red \_

Roses are red violets \_

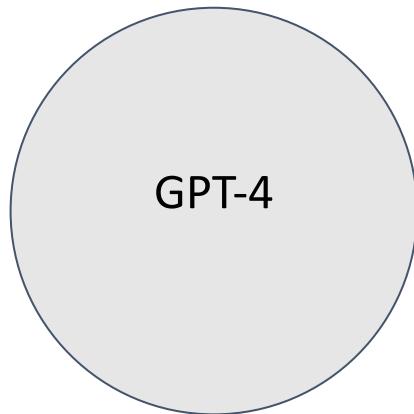
Roses are red violets are \_

Roses are red violets are blue \_

# GPT-4



Text input  
(prompt)  
OR  
Text input  
and image



Text output

# College level exams

- GPT-4 matching human level performance at some exams
- Model performance progressing quickly

# College level exams

GPT-4 technical report: <https://arxiv.org/pdf/2303.08774.pdf>

# MMLU – Massive Multitask Language Understanding

- Multiple-choice questions in 57 subjects
- Includes STEM, humanities and the social sciences.

GPT-4 technical report: <https://arxiv.org/pdf/2303.08774.pdf>



# MMLU: Physics

When you drop a ball from rest it accelerates downward at  $9.8 \text{ m/s}^2$ . If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is

- (A)  $9.8 \text{ m/s}^2$
- (B) more than  $9.8 \text{ m/s}^2$
- (C) less than  $9.8 \text{ m/s}^2$
- (D) Cannot say unless the speed of throw is given.



# MMLU: Physics

When you drop a ball from rest it accelerates downward at  $9.8 \text{ m/s}^2$ . If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is

- (A)  $9.8 \text{ m/s}^2$
- (B) more than  $9.8 \text{ m/s}^2$
- (C) less than  $9.8 \text{ m/s}^2$
- (D) Cannot say unless the speed of throw is given.

✓

✗

✗

✗



# MMLU: Microeconomics

One of the reasons that the government discourages and regulates monopolies is that

- (A) producer surplus is lost and consumer surplus is gained.
- (B) monopoly prices ensure productive efficiency but cost society allocative efficiency.
- (C) monopoly firms do not engage in significant research and development.
- (D) consumer surplus is lost with higher prices and lower levels of output.



# MMLU: Microeconomics

One of the reasons that the government discourages and regulates monopolies is that

- (A) producer surplus is lost and consumer surplus is gained. ✗
- (B) monopoly prices ensure productive efficiency but cost society allocative efficiency. ✗
- (C) monopoly firms do not engage in significant research and development. ✗
- (D) consumer surplus is lost with higher prices and lower levels of output. ✓



# MMLU: Medicine

A 33-year-old man undergoes a radical thyroidectomy for thyroid cancer. During the operation, moderate hemorrhaging requires ligation of several vessels in the left side of the neck. Postoperatively, serum studies show a calcium concentration of 7.5 mg/dL, albumin concentration of 4 g/dL, and parathyroid hormone concentration of 200 pg/mL. Damage to which of the following vessels caused the findings in this patient?

- (A) Branch of the costocervical trunk
- (B) Branch of the external carotid artery
- (C) Branch of the thyrocervical trunk
- (D) Tributary of the internal jugular vein



# MMLU: Medicine

A 33-year-old man undergoes a radical thyroidectomy for thyroid cancer. During the operation, moderate hemorrhaging requires ligation of several vessels in the left side of the neck.

Postoperatively, serum studies show a calcium concentration of 7.5 mg/dL, albumin concentration of 4 g/dL, and parathyroid hormone concentration of 200 pg/mL. Damage to which of the following vessels caused the findings in this patient?

- (A) Branch of the costocervical trunk ✗
- (B) Branch of the external carotid artery ✗
- (C) Branch of the thyrocervical trunk ✓
- (D) Tributary of the internal jugular vein ✗

Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar

Source: GPT-4 Technical Report

# Why GPT-4?

# Objectives of GPT-3

- Predict the next word



# Challenges with GPT-3

- Doesn't follow user instructions
- Can generate toxic language
- Can make up facts

# Objectives of GPT-3.5 / GPT-4

- Helpful and able to follow instructions
- Not toxic – for example, hateful speech, foul language
- Less likely to fabricate information or hallucinate

# OpenAI playground

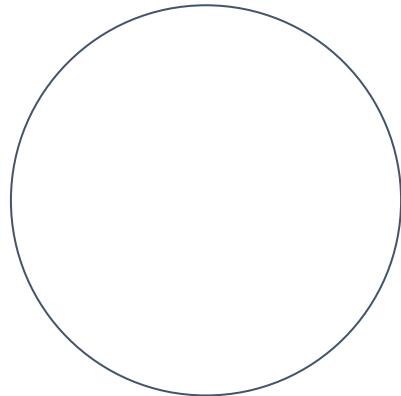
Create a shopping list

# Comparing GPT-4 to GPT-3 and GPT-3.5

# GPT-3?

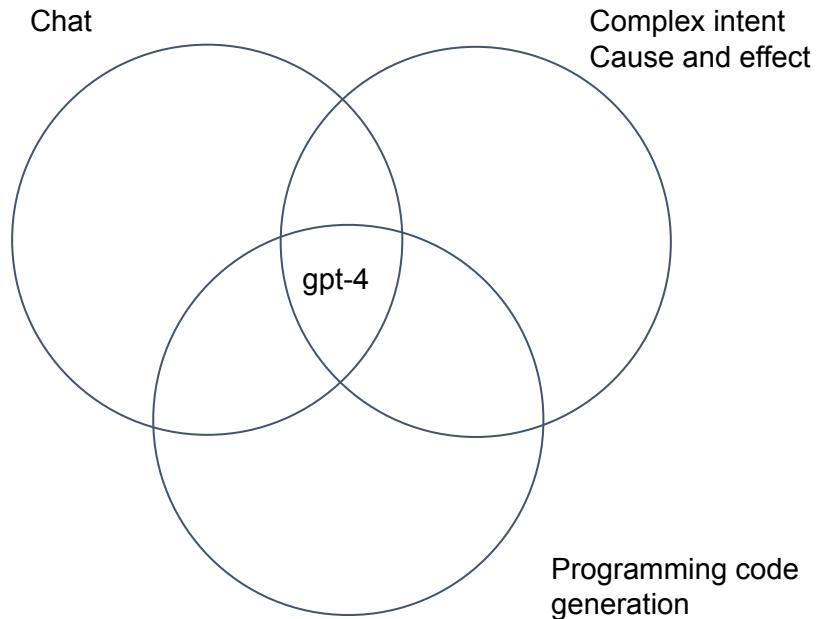
- davinci

Text generation

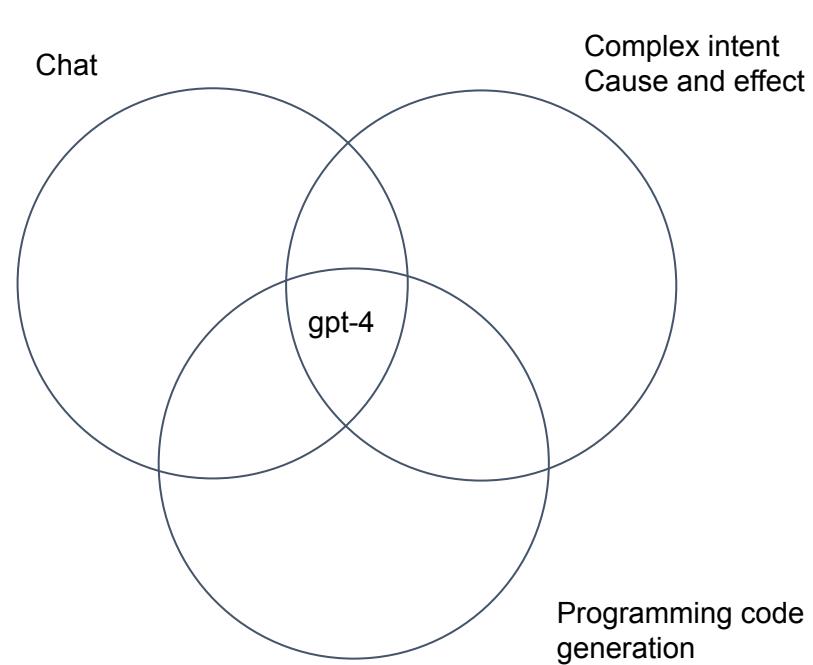
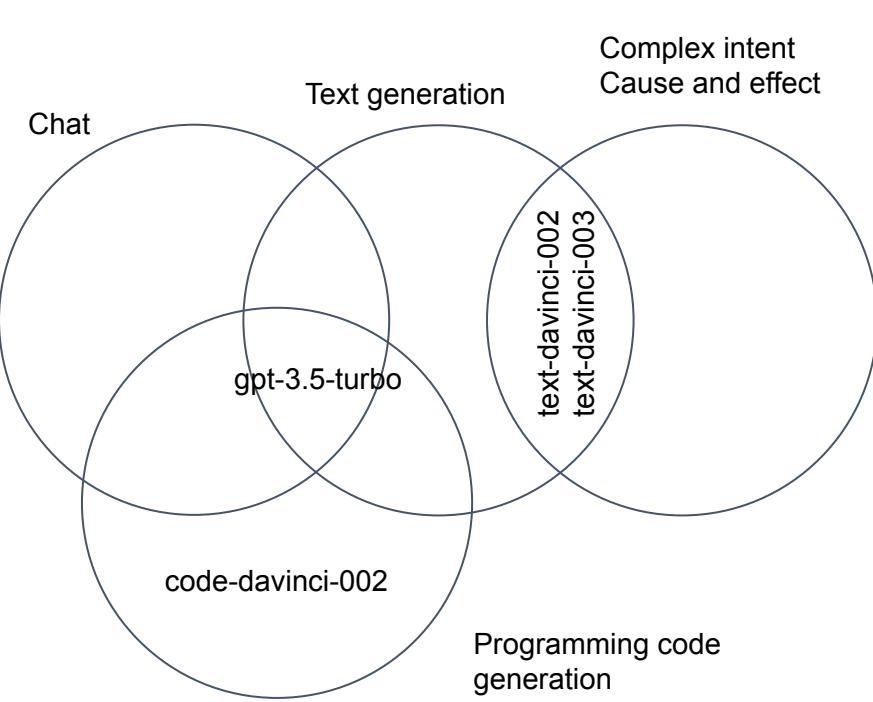


# GPT-4

- gpt-4
- Gpt-4-32k
- gpt-4-turbo



# What Is GPT-4?



# Differences between the GPT-4 models

Prompt



# Differences between the GPT-4 models

Prompt



Completion



# Differences between the GPT-4 models

Prompt



Completion



Context window

# Differences between the GPT-4 models

**gpt-4:** 8,000 tokens

**gpt-4-32k:** 32,000 tokens

# Differences between the GPT-4 models

**gpt-4:** 8,000 tokens

**gpt-4-32k:** 32,000 tokens

**gpt-3:** 2,000 tokens

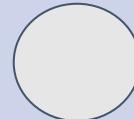
**gpt-3.5:** 4,000 tokens

**claude-3?**

# Pricing

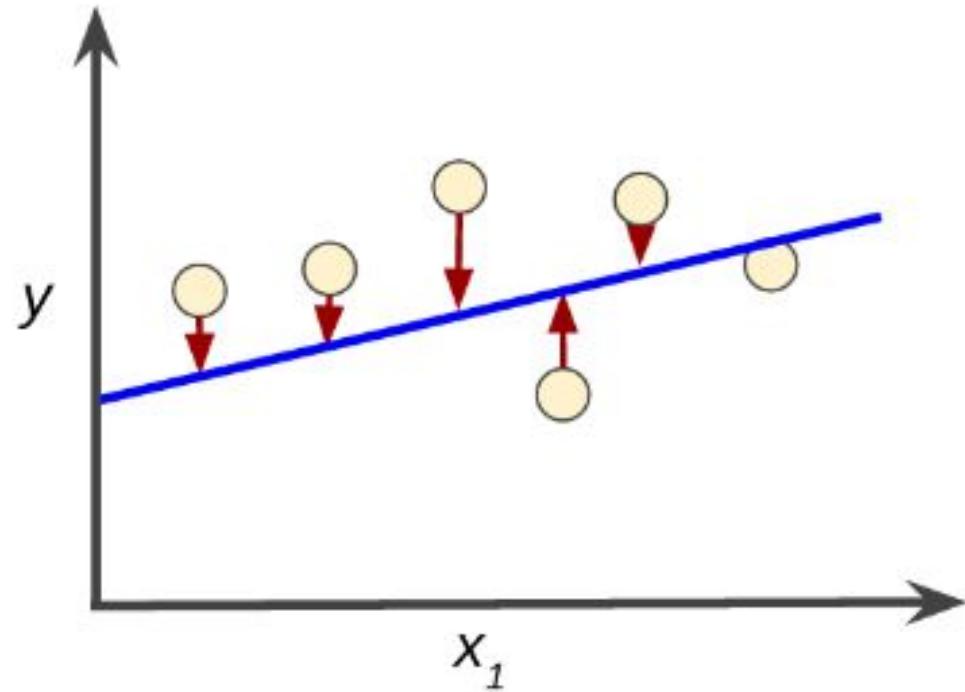
Model	PROMPT Price per 1,000 Tokens (About 750 words)	COMPLETION Price per 1,000 Tokens (About 750 words)
text-davinci-002	\$0.02	\$0.02
text-davinci-003		
code-davinci-002		
gpt-3.5-turbo	\$0.002	\$0.002
gpt-4 (8K context)	\$0.03	\$0.06
gpt-4 (32K context)	\$0.06	\$0.12

# Model Architecture Differences

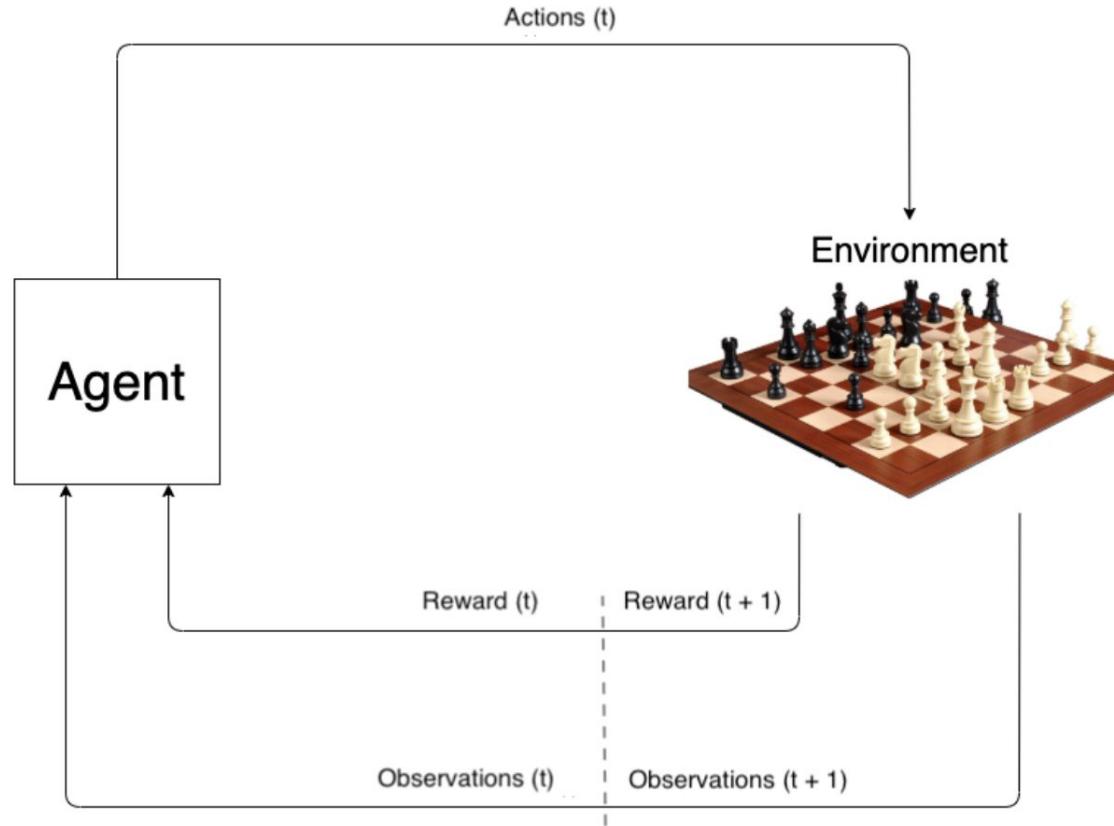
Model	COMPLETIONS: Price per 1,000 Tokens (About 750 words)	Model Size (Number of Parameters)
text-davinci-002 text-davinci-003 code-davinci-002	\$0.02	
gpt-3.5-turbo	\$0.002	
gpt-4	\$0.06/\$0.12	Unknown

# How Was GPT-4 Trained?

# How ML models are trained



# Reinforcement Learning





# Reinforcement Learning from Human Feedback

Step 1

## Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

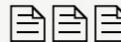
Explain the moon landing to a 6 year old

A labeler demonstrates the desired output behavior.

Some people went to the moon...

This data is used to fine-tune GPT-3 with supervised learning.

SFT



Training language models to follow instructions with human feedback - Long et al (Open AI)

# Example: Summarize a News Article



[Submit](#)[Skip](#)

«

Page  / 11

»

Total time: 05:39

**Instruction**

Summarize the following news article:

====

{article}

====

[Include output](#)**Output A**

summary1

**Rating (1 = worst, 7 = best)**

1

2

3

4

5

6

7

Fails to follow the correct instruction / task ?  Yes  No

Inappropriate for customer assistant ?  Yes  No

Contains sexual content  Yes  No

Contains violent content  Yes  No

Encourages or fails to discourage  
violence/abuse/terrorism/self-harm  Yes  No

Denigrates a protected class  Yes  No

Gives harmful advice ?  Yes  No

Expresses moral judgment  Yes  No

**Notes**

(Optional) notes

## Ranking outputs

### To be ranked

**B** A team of researchers from Yale University and University of California, Davis studied the vocalization patterns of several different types of parrots. They found that parrots like to mimic human speech, and can produce a wide range of sounds, such as whistles, squawks, and other types of vocalizations...

#### Rank 1 (best)

**A** A research group in the United States has found that parrots can imitate human speech with ease, and some of them can even do so in the same way as humans. This group studied the sounds that parrots make in their natural habitats and found that they use their tongues and beaks in ways that are strikingly...

**C** Parrots have been found to have the ability to understand numbers. Researchers have found that parrots can understand numbers up to six. In a series of experiments, the parrots were able to identify the amount of food items under a number of cups...

#### Rank 2

#### Rank 3

**E** Scientists have found that green-winged parrots can tell the difference between two noises that are the same except for the order in which they are heard. This is important because green-winged parrots are known to imitate sounds. This research shows that they are able to understand the difference between sounds.

**D** Current research suggests that parrots see and hear things in a different way than humans do. While humans see a rainbow of colors, parrots only see shades of red and green. Parrots can also see ultraviolet light, which is invisible to humans. Many birds have this ability to see ultraviolet light, an ability

#### Rank 4

#### Rank 5 (worst)

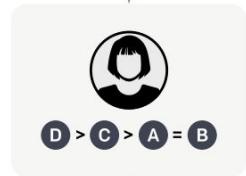
Step 2

## Collect comparison data, and train a reward model.

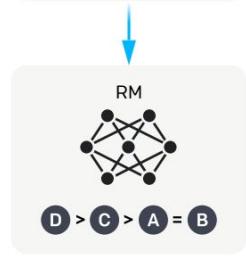
A prompt and  
several model  
outputs are  
sampled.



A labeler ranks  
the outputs from  
best to worst.



This data is used  
to train our  
reward model.



Training language models to follow instructions with human feedback - Long et al (Open AI)

Step 3

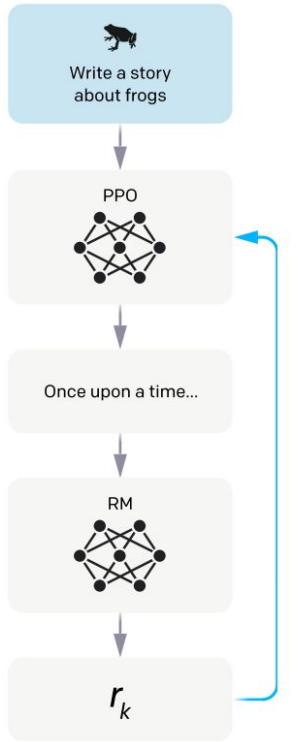
### Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

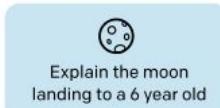


Training language models to follow instructions with human feedback - Long et al (Open AI)

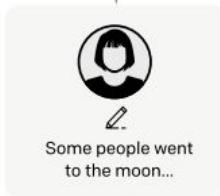
Step 1

**Collect demonstration data, and train a supervised policy.**

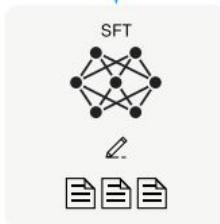
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3 with supervised learning.



Step 2

**Collect comparison data, and train a reward model.**

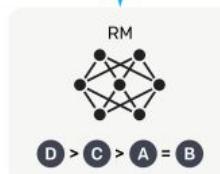
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

**Optimize a policy against the reward model using reinforcement learning.**

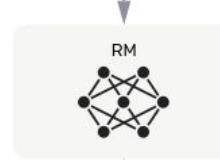
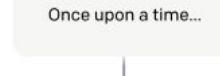
A new prompt is sampled from the dataset.



The policy generates an output.

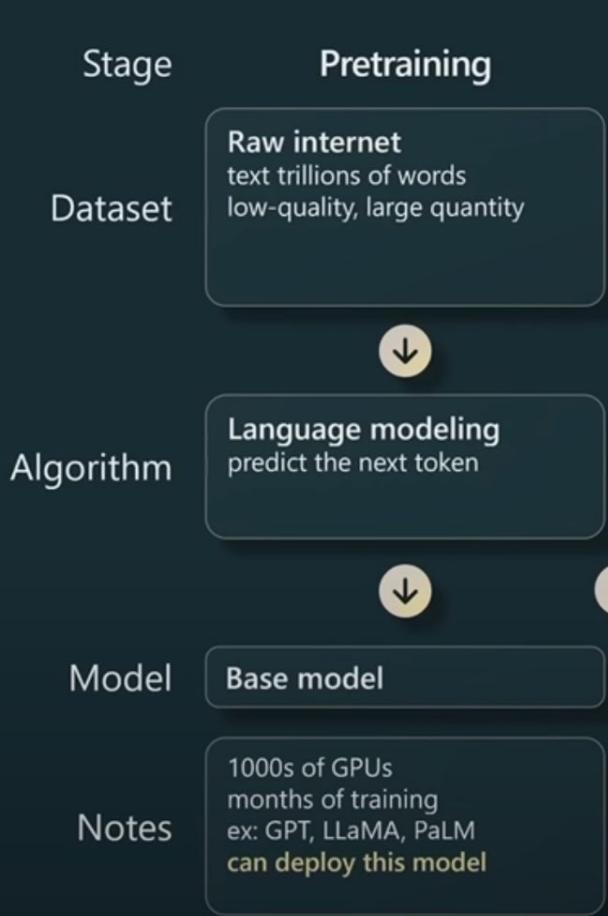


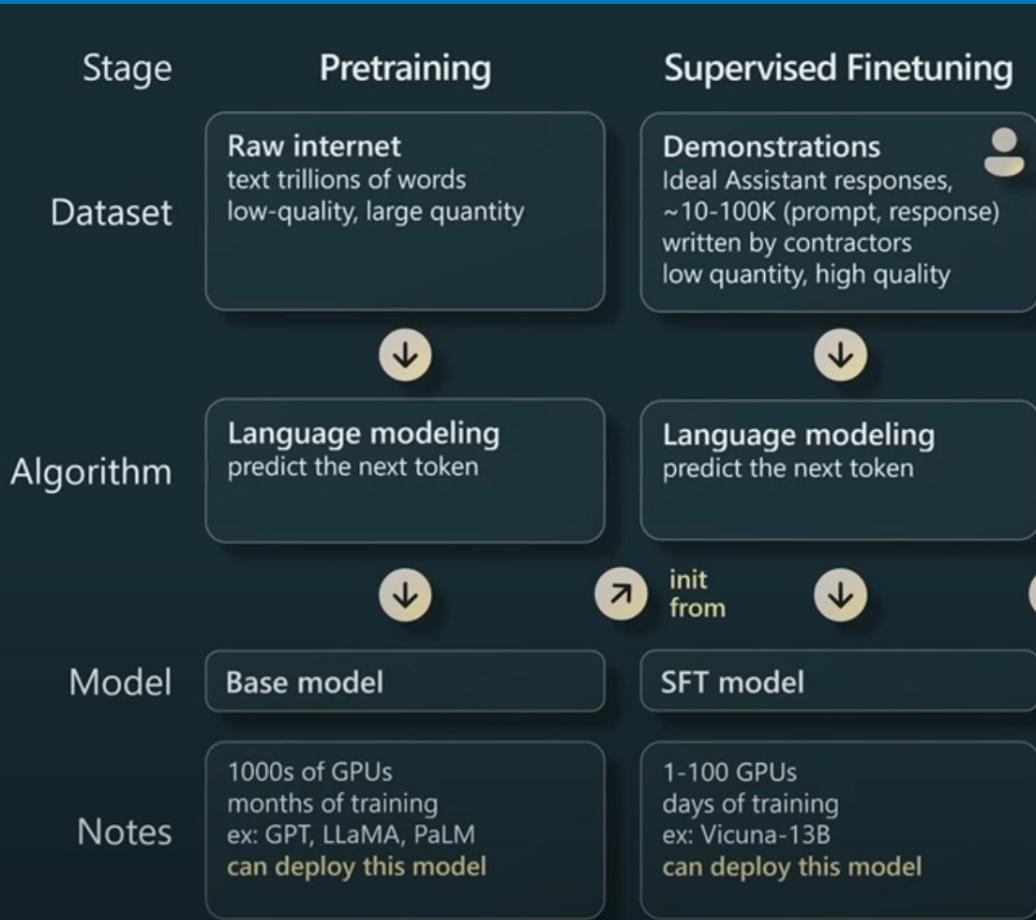
The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.

and 1.5 years later ...





Stage	Pretraining	Supervised Finetuning	Reward Modeling	Reinforcement Learning
Dataset	Raw internet text trillions of words low-quality, large quantity	Demonstrations Ideal Assistant responses, ~10-100K (prompt, response) written by contractors low quantity, high quality	Comparisons 100K –1M comparisons written by contractors low quantity, high quality	Prompts ~10K-100K prompts written by contractors low quantity, high quality
Algorithm	Language modeling predict the next token	Language modeling predict the next token	Binary classification predict rewards consistent w preferences	Reinforcement Learning generate tokens that maximize the reward
Model	Base model	SFT model	RM model	RL model
Notes	1000s of GPUs months of training ex: GPT, LLaMA, PaLM <b>can deploy this model</b>	1-100 GPUs days of training ex: Vicuna-13B <b>can deploy this model</b>	1-100 GPUs days of training	1-100 GPUs days of training ex: ChatGPT, Claude <b>can deploy this model</b>

# Limitations of GPT-4

# Model Weaknesses

- Only recently been able to fine-tune (Not possible on release)
- Doesn't update its knowledge in real-time.
- Makes up facts.

# Lawyer apologizes for fake court citations from ChatGPT

By Ramishah Maruf, CNN

Updated 3:28 PM EDT, Sun May 28, 2023



# Why fine-tuning?

*The Journal of Infectious Diseases*

EDITORIAL COMMENTARY



## Pneumococcal Carriage and Seroepidemiology Studies to Measure Current and Future Pneumococcal Conjugate Vaccine Effectiveness

Keith P. Klugman and Gail L. Rodgers

Pneumonia, Surveillance and Epidemic Control Programs, Bill & Melinda Gates Foundation, Seattle, Washington, USA

Two current questions in pneumococcal conjugate vaccine (PCV) development for children are whether immunization schedules are optimal to maintain direct and indirect protection of licensed vaccines, and what are the potential roles of next-generation higher-valent vaccines to further reduce pediatric pneumococcal disease.

invasive pneumococcal disease (IPD) due to the vaccine serotypes, but this measure has been frustrated by the coronavirus disease 2019 (COVID-19) pandemic, which has reduced the incidence of IPD in the United Kingdom and many other countries, largely as a result of masking and restrictions on indoor

children is that potentially key invasive serotypes, such as serotypes 8 and 12F, which have dominated adult disease after PCV13 introduction in children in the United Kingdom, were only rarely detected in carriage in this study. So while carriage among healthy children is useful to describe the potential distribution of

# GPT-4 - hallucination

These are the topics that I am covering in an O'reilly online training on GPT-4. Write a catchy introduction.

What is GPT-4

Why GPT-4?

Comparing GPT-4 to GPT-3 and GPT-3.5

How was GPT-4 trained?

What are the limitations of GPT-4?

HELM

# Colab notebook: GPT exercises

[https://colab.research.google.com/drive/1-rSgn9pM5zyRXrxchzcxGZ\\_Wm\\_FTuo1K?usp=sharing](https://colab.research.google.com/drive/1-rSgn9pM5zyRXrxchzcxGZ_Wm_FTuo1K?usp=sharing)

g

<https://mistral.ai/news/mixtral-8x22b/>

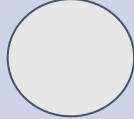
# Pricing

Model	PROMPT Price per 1,000 Tokens (About 750 words)	COMPLETION Price per 1,000 Tokens (About 750 words)
GPT-3.5 Turbo 4K	\$0.0015	\$0.002
GPT-3.5 Turbo 16K	\$0.001	\$0.002
GPT-3.5 Turbo 4K fine-tuning	\$0.012 (Training \$0.008)	\$0.003
GPT-3.5 Turbo 16K fine-tuning	\$0.003 (Training \$0.008)	\$0.006

# Pricing

Model	PROMPT Price per 1,000 Tokens (About 750 words)	COMPLETION Price per 1,000 Tokens (About 750 words)
GPT-4 Turbo 8K	\$0.03	\$0.06
GPT-4 Turbo 16K	\$0.06	\$0.12
GPT-4 Turbo 128K	\$0.01	\$0.03

# Model Architecture Differences

Model	Model Size (Number of Parameters)
text-davinci-002	
text-davinci-003	
code-davinci-002	
gpt-3.5-turbo	
gpt-4	Unknown
gpt-4-turbo	Unknown



# Transformers in production





curling objective





# BERT

Bidirectional Encoder Representations from  
Transformers



# Where are Transformers used in production?

what's the main objective for curling in the olympics

X |

All Images News Videos Shopping More Tools

About 18,900,000 results (0.65 seconds)

The goal for each team is **to get stones as close to the center of the house as possible and earn points based on the positioning of their stones.** Only one team can score in an end, and points are only awarded if the stones are touching the house. The team with the most points after 10 ends is the winner. 14 Feb 2022

<https://www.sportingnews.com> › olympics › news › curlin...

How does curling work? Explaining the rules and scoring for ...



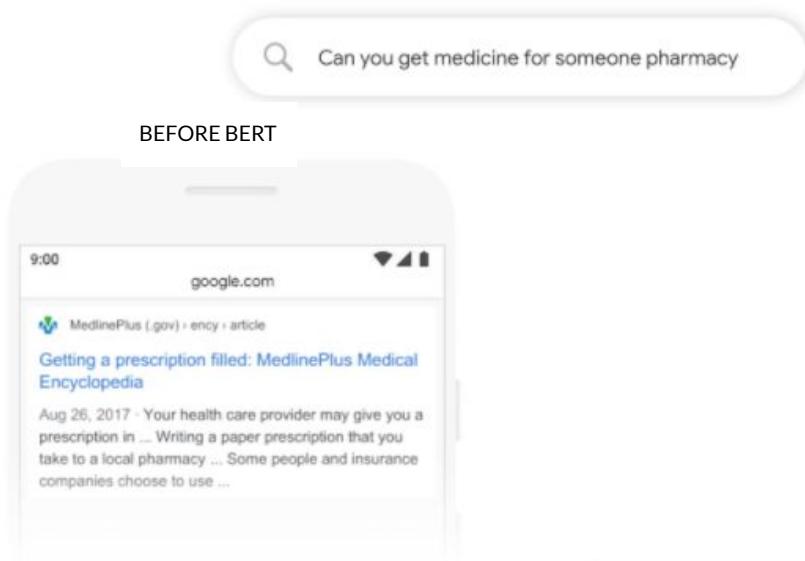
About featured snippets



Feedback



# Transformers in production





# Transformers in production

Can you get medicine for someone pharmacy

BEFORE BERT

google.com

MedlinePlus (.gov) · ency · article

Getting a prescription filled: MedlinePlus Medical Encyclopedia

Aug 26, 2017 · Your health care provider may give you a prescription in ... Writing a paper prescription that you take to a local pharmacy ... Some people and insurance companies choose to use ...

AFTER BERT

google.com

HHS.gov · hipaa · for-professionals

Can a patient have a friend or family member pick up a prescription ...

Dec 19, 2002 · A pharmacist may use professional judgment and experience with common practice to ... the patient's best interest in allowing a person, other than the patient, to pick up a prescription.



# What was BERT trained on?

BERT - Wikipedia and BooksCorpus (11,000 unpublished books)



# What tasks was BERT trained?

- Masked Language Model (MLM)
- Next Sentence Prediction (NSP)



The Tokyo Olympic games were <masked> from 2020 to 2021.



# Masked Language Modelling (MLM)

The Tokyo Olympic games were <masked> from 2020 to 2021.



# Masked Language Modelling (MLM)

The Tokyo Olympic games were postponed from 2020 to 2021.



# Next sentence prediction (NSP)

The Tokyo Olympic games were postponed from 2020 to 2021. This is the first instance in the history of the Olympics as previous games had been cancelled but not rescheduled.



# Why MLM and NSP?

**BERT gets a good understanding of English language.**



# Pre-training: BERT

	BERT
Year	2018
Number of parameters	109M
Training time	12 days
Infrastructure	8 x V100 GPUs (*)
Size of dataset used for training	16GB
Training tokens (dataset)	250B
Dataset source	Wikipedia
	Book corpus



# What are tokens?

1500 words is approximately equivalent to 2400 tokens



# What are tokens?

1500 words is approximately equivalent to 2400 tokens

A word is approximately 1.4 tokens



# What are tokens?

1500 words is approximately equivalent to 2400 tokens

A word is approximately 1.4 tokens

A novel is 100,000 words, or 140,000 tokens



# What are tokens?

BERT was trained on 250B  
tokens or:  
1.8 million novels

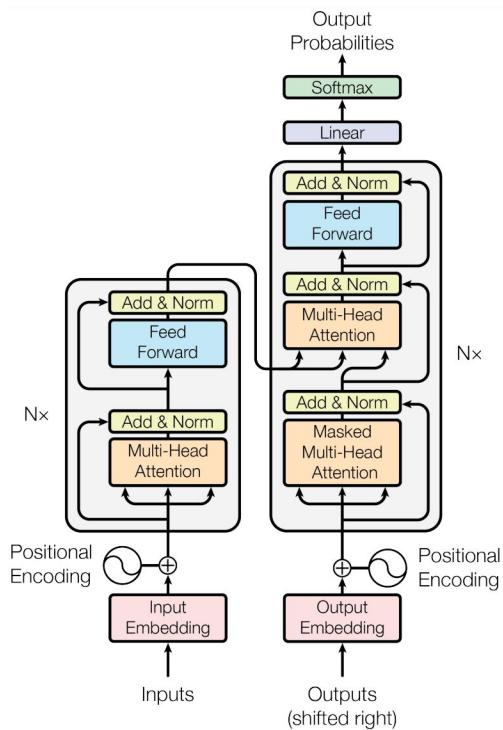




# Transformer: Architecture Overview

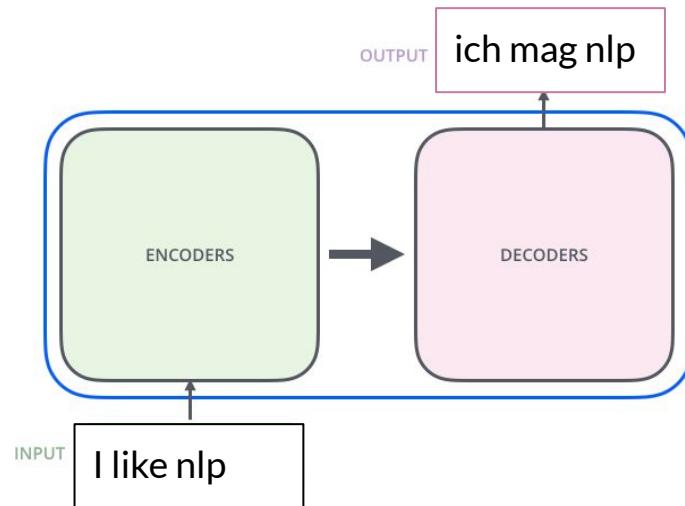


# Transformer architecture



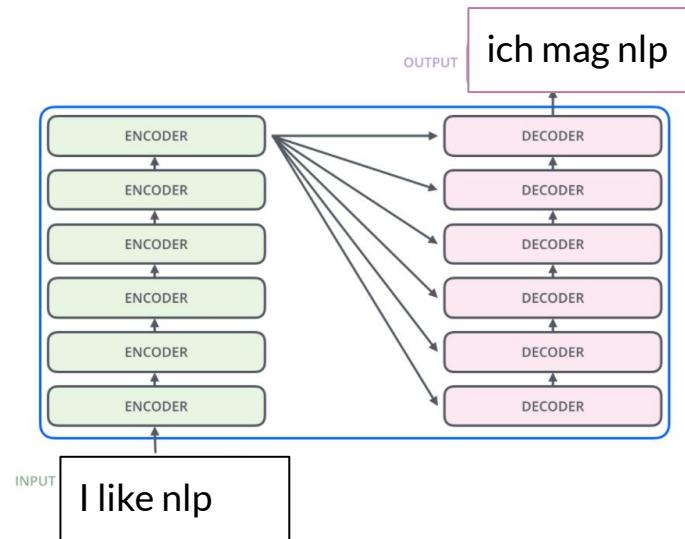


# Transformer overview





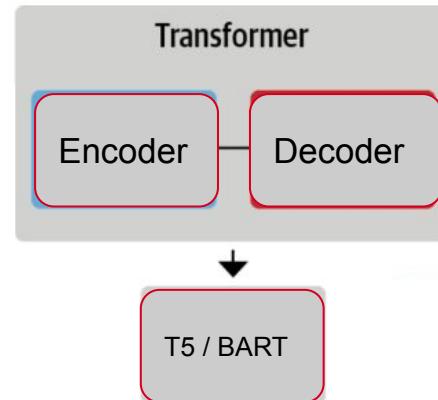
# Transformer overview





# Encoder-decoder model

- Generative tasks
- BART
- T5





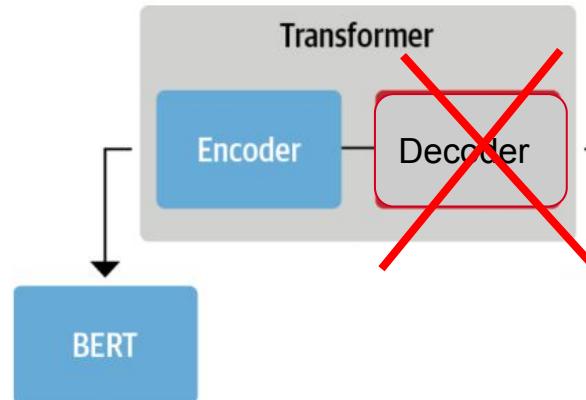
# Encoder-only model

Understanding of input

- Sentence classification
- Named Entity Recognition

Family of BERT models:

- BERT, RoBERTa, DistilBERT ...



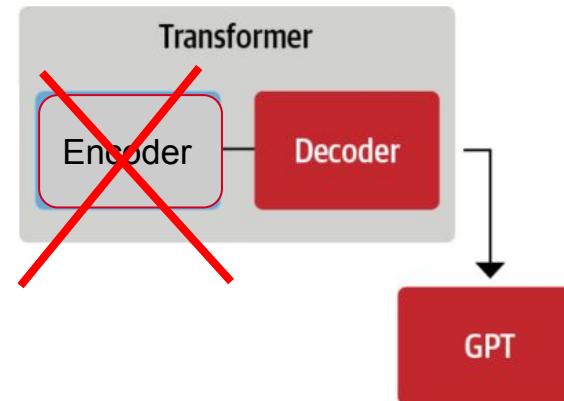


# Decoder-only model

- Generative tasks

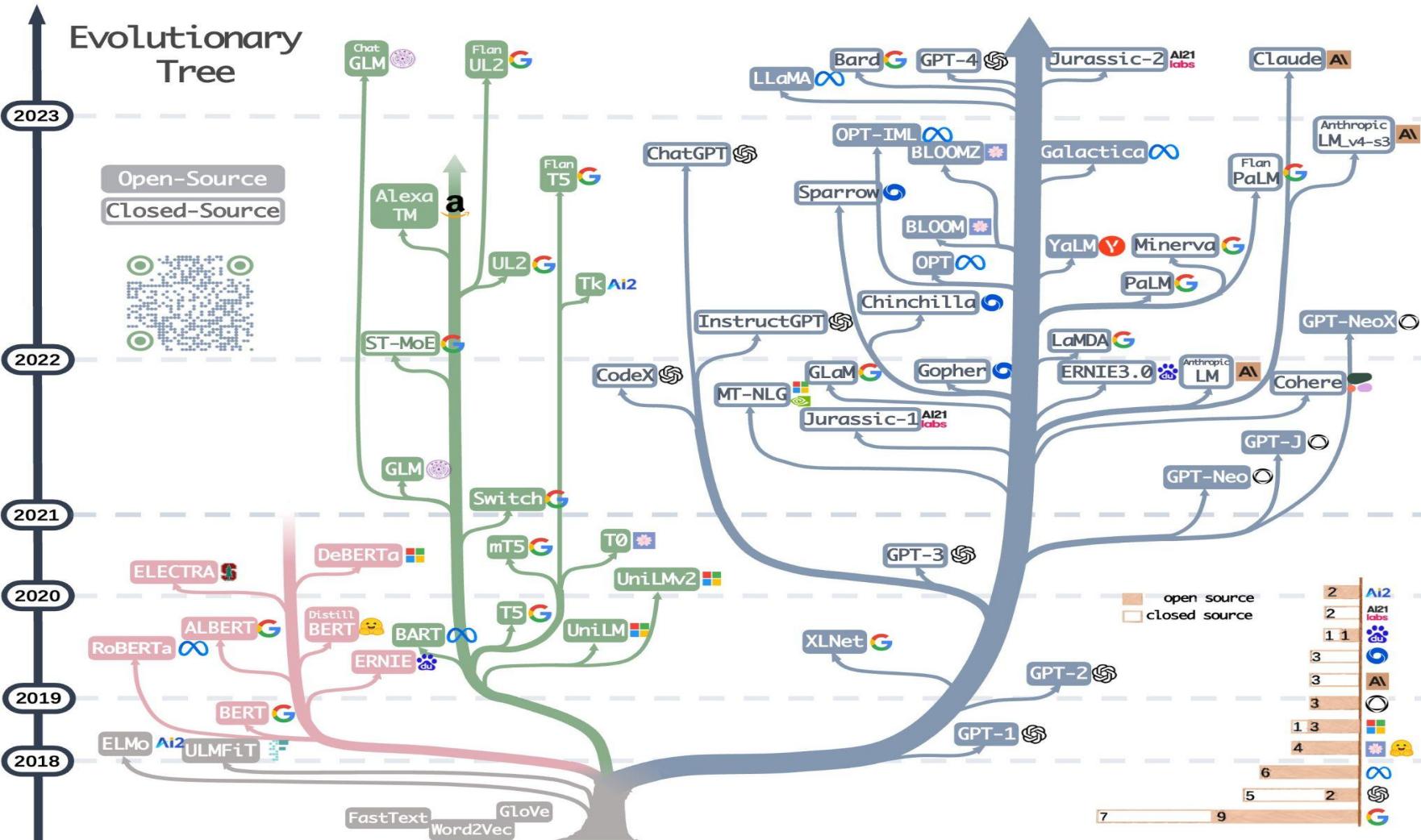
Examples:

- GPT
- GPT-2
- GPT-3





<https://github.com/Mooler0410/LLMsPracticalGuide>

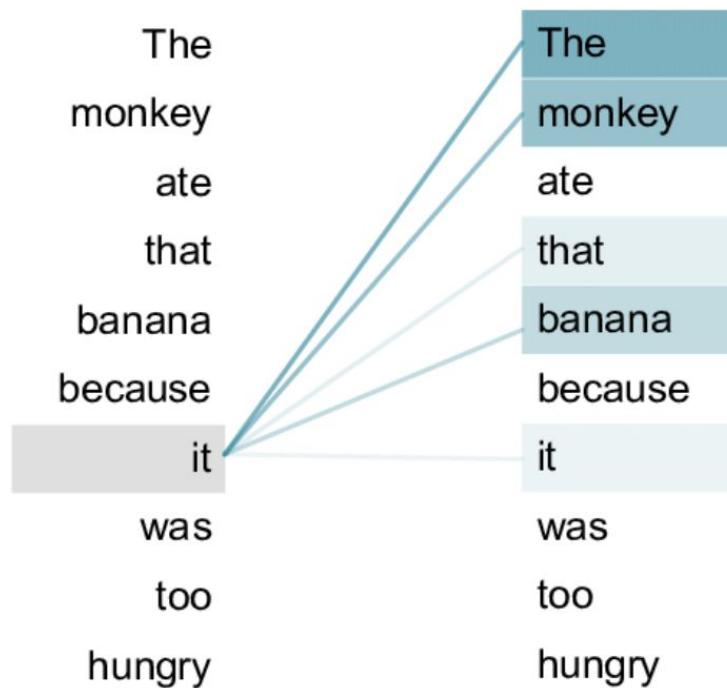


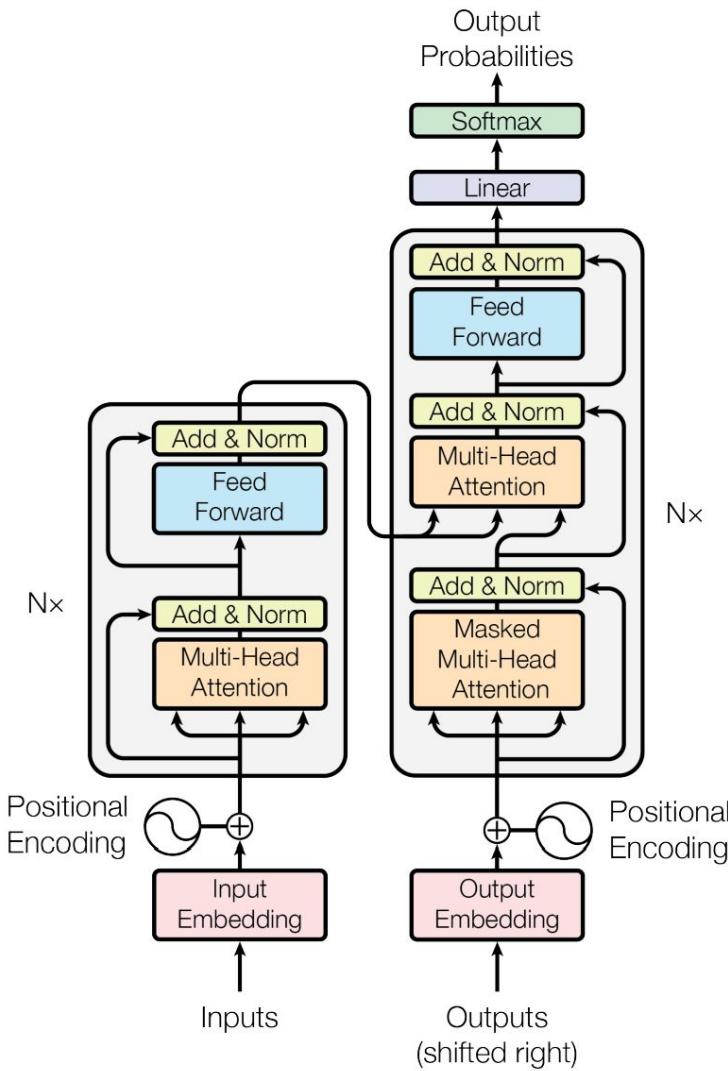


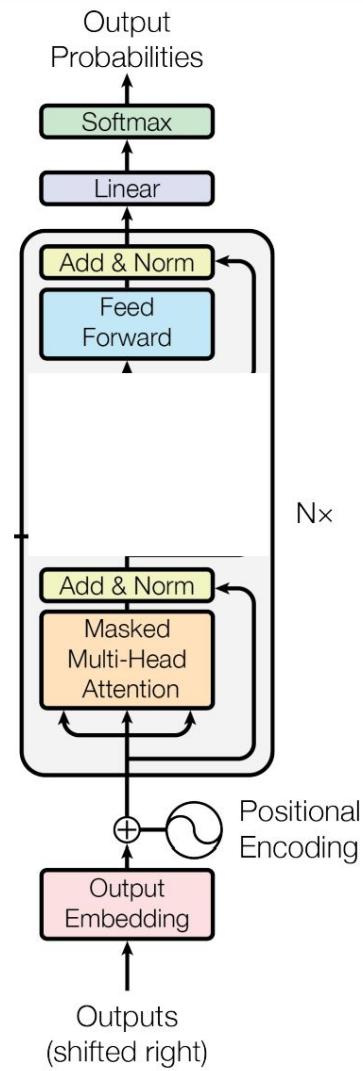
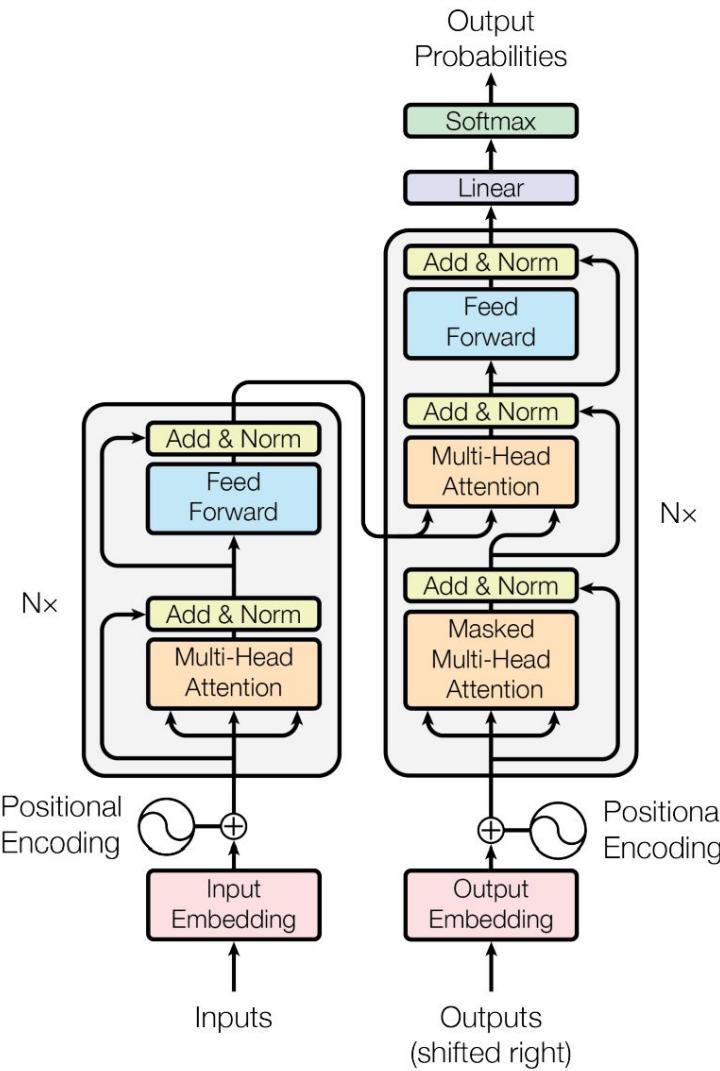
# Self-attention



# self-attention









# self-attention

Attention has 3 inputs:

- Q (query)
- K (key)
- V (value)



# self-attention

Attention(**Q, K, V**)



# self-attention

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{n}}$$



# self-attention

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{n}}\right)$$



# self-attention

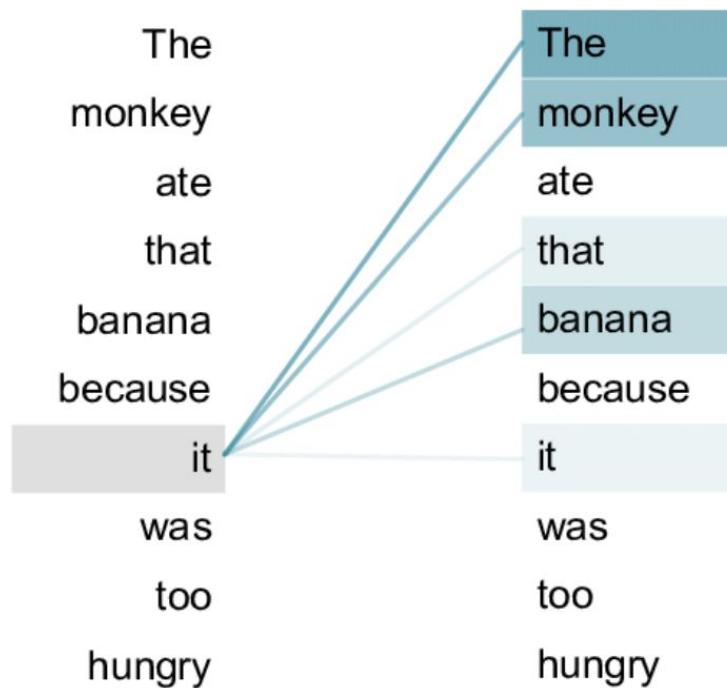
$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{n}}\right)\mathbf{V}$$



# Multi-headed attention

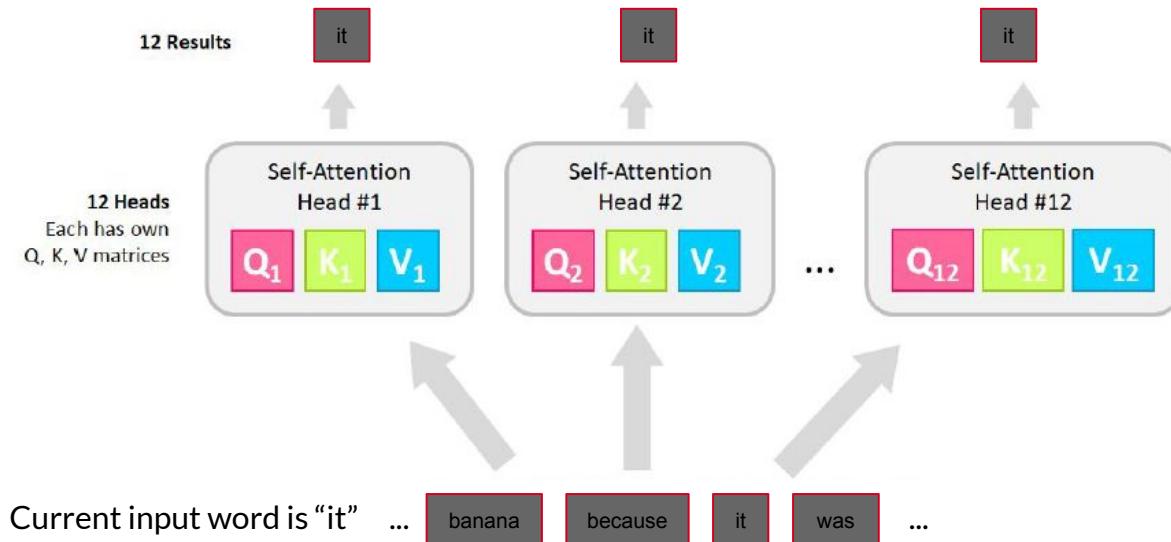


# self-attention



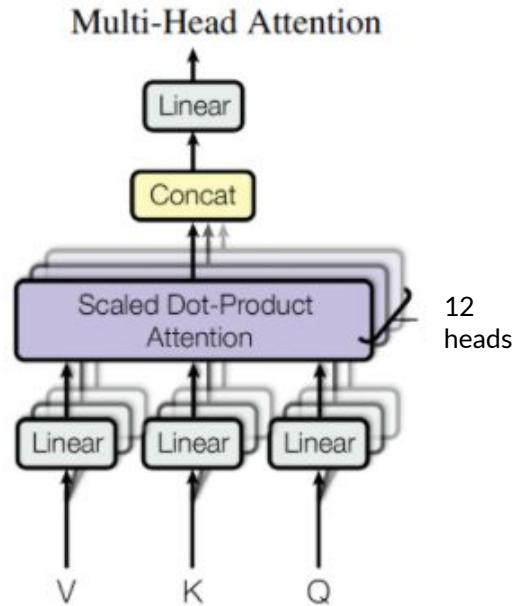


# Multi-headed attention





# Multi-Head Attention





# Week 1 (today)

## GPT-3 and why we needed GPT-4

- Where GenAI stands now
- GPT-3+ and GPT-4
- Limitations of GPT-4
- RLHF
- Q&A
- Break

## Transfer learning and fine-tuning

- BERT
- Transfer learning and fine-tuning
- Q&A
- Break

## Transformer architecture overview

- Encoders and decoders
- Q&A
- Break

The background features a vibrant red-to-yellow gradient. Overlaid on this gradient are three large, semi-transparent white circles of varying sizes, creating a sense of depth and motion.

O'REILLY®